# Evaluation of Medical Language Models

Aman Sinha

UNIVERSITÉ DE LORRAINE
34 COURS LÉOPOLD — CS 25233
54052 NANCY CEDEX
TÉL : 03 54 50 54 00
CONTACT@UNIV-LORRAINE.FR
WWW.UNIV-LORRAINE.FR

Ecole doctorale IAEM Lorraine

**Université de Lorraine - ICANS**
Institut Elie Cartan de Lorraine (UMR 7502)
Analyse et Traitement Informatique de la Langue Française (UMR 7118)
Institut de cancérologie Strasbourg Europe (ICANS)

# Evaluation of Medical Language Models

A dissertation submitted to the Université de Lorraine

by Aman SINHA

in partial fulfillment of the requirements for the degree of
Docteur en Informatique de l'Université de Lorraine

December 12th, 2025

***Jury Members***

| | |
|---|---|
| *Supervisor* | Marianne CLAUSEL, Université de Lorraine |
| *Co-supervisor* | Mathieu CONSTANT, Université de Lorraine |
| *Reviewers* | Lucie FLEK, University of Bonn |
| | Douglas TEODORO, Université de Genève |
| *Examiners* | Maxime AMBLARD, Université de Lorraine, *Jury President* |
| | Aurélie NÉVÉOL, CNRS / Université Paris-Saclay |
| | Eric VILLEMONTE DE LA CLERGERIE, INRIA |
| *Guest Member* | Xavier COUBEZ, Institut Strauss |

# Acknowledgments

Firstly, I would like to express my sincere gratitude to Marianne Clausel, Mathieu Constant, and Xavier Coubez, who have been strong pillars of support throughout my research journey in France. I am grateful to Marianne for her perpetual enthusiasm for exploring research ideas and also for my first French coffee experience in 2020, showing me how café fuels mathematicians and computer scientists. I would like to thank Mathieu for generously sharing his expertise in NLP and research experiences, and for making sure that while exploring, I also remain grounded. I would also like to thank Xavier Coubez for all the elaborate and encouraging feedback from the day we met virtually and all throughout the different research endeavors I brought to him.

I am extremely thankful to my jury members, Lucie Flek, Douglas Teodoro, Aurélie Névéol, Maxime Amblard, and Eric de la Clergerie, for taking their time to provide their valuable feedback and expert validation on my work.

I would like to express my thanks to Jörg Tiedmann and Helsinki-NLP group for hosting me for my research visit at the University of Helsinki, enabling me to meet several cool researchers and kind people.

A lot of people have contributed to my PhD journey over the past four years in so many ways. I am thankful to Satyam for helping me recover all my thesis work after my laptop crashed at one point. I am grateful to Ioana for her support during the last two years of my PhD for all our projects, and for the many discussions, both scientific and otherwise, that helped me to get back on my feet to work when I was having a difficult time. I am very thankful to Priyansh, Pallu, Vaidya, Shugau, Bhdani, Anshu, and Par for supporting me through the many breakdowns I had. To Rauq, thank you for your kindness since the first day we met in Helsinki and for checking on me randomly during the writing period. And to Peppa, thank you for your patience during all my venting sessions.

In my family, a special mention to my sister Nishu, my strongest supporter, who, despite being younger, never failed to inspire me with her cheerfulness and attitude towards everything. To my brother Shubh, thank you for staying so strong while I was so far away. I am thankful to Ishu, Ashu, and Lalla for all the game nights. I am very thankful to Shalu for taking care of Nishu and helping her to understand how jokes work. And to my dad, for all the discomfort he endured throughout his life and for still standing every day to support my career and comfort (these words are not enough).

Another special mention to Rohit, who stood by me through all the highs and lows as a very dear friend, taking on roles ranging from research comrade to moral compass, thank you! Lastly, Timothée – thank you for existing; and thank you again, Mathieu, for introducing him to me. What began with Timothée as my research senpai eventually became 'Timothée-san', followed by our various research and acronym adventures that have made my PhD journey significantly memorable.

Finally, I would like to thank Université de Lorraine, ICANS Strasbourg, and the

IAEM Doctoral School for their continuous support throughout my PhD and for the DREAM mobility grant. I am also grateful to Madame Nathalie Benito for her patience and invaluable help with all travel and administrative arrangements.

*While there are no words to fully express my gratitude, I will always be indebted to you. This thesis is a small token that I dedicate to you, Ma.*

# Abstract

Medical language is inherently complex, as it constitutes a specialized sublanguage characterized by unstructured text, domain-specific vocabulary, and considerable heterogeneity, distinguishing it from general language. Understanding medical language requires expert knowledge, which is typically limited to clinicians and medical professionals. Similarly, deep learning models trained on general-domain datasets often fail to transfer their performance effectively to the medical domain. In this thesis, we focus on two primary reasons for this limitation: the intrinsic nature of medical datasets and the lack of domain-specific expertise embedded in the models themselves.

To systematically investigate the challenges arising from the *nature of medical data*, this thesis proposes a taxonomy categorizing medical data into three types based on their source: social media (SOC), clinical records (CLIN), and scientific literature (SCI). Each source exhibits distinct linguistic styles, degrees of structure, and information content, necessitating tailored modeling approaches. Building on this taxonomy, we conduct a comprehensive study of medical data across multiple levels of linguistic complexity — token-level, sentence-level, document-level, and temporal-level — examining the limitations of language models when trained on single-source data.

The second contribution of this thesis addresses the challenges identified in single-sourced models by exploring *multi-sourced medical language models*. This approach leverages additional sources of information, including external knowledge bases, knowledge graphs, and large language models, to complement single-source models and mitigate domain-specific limitations.

The third contribution is a reflective investigation into the *proficiency gap between language models and human experts*. This analysis considers two dimensions: (a) model-centric issues, focusing on the characteristics of language models in the medical domain, their design choices, and the trade-offs that arise during model development and deployment; and (b) data-centric issues, examining differences in concept understanding, knowledge alignment, and task difficulty perception between language models and human experts.

Overall, this thesis adopts a critical perspective on medical language models, providing insights into what is required to build an "artificial health expert" and highlighting substantial opportunities for future research in this direction.

# Resumé en Français

## Résumé court

Le langage médical est intrinsèquement complexe, car il constitue un sous-langage spécialisé caractérisé par un texte non structuré, un vocabulaire spécifique au domaine et une hétérogénéité considérable, ce qui le distingue du langage général. La compréhension du langage médical nécessite des connaissances spécialisées, qui sont généralement limitées aux cliniciens et aux professionnels de la santé. De même, les modèles d'apprentissage profond entrainés sur des ensembles de données du domaine général ne parviennent souvent pas à transférer efficacement leurs performances au domaine médical. Dans cette thèse, nous nous concentrons sur deux raisons principales à cette limitation : la nature intrinsèque des ensembles de données médicales et le manque d'expertise spécifique au domaine intégrée dans les modèles eux-mêmes.

Afin d'étudier de manière systématique les défis découlant de la *nature des données médicales*, cette thèse propose une taxonomie classant les données médicales en trois types en fonction de leur source: les réseaux sociaux (SOC), les dossiers cliniques (CLIN) et la littérature scientifique (SCI). Chaque source présente des styles linguistiques, des degrés de structure et des contenus informationnels distincts, qui nécessitent des approches de modélisation adaptées. En nous appuyant sur cette taxonomie, nous menons une étude approfondie des données médicales à plusieurs niveaux de complexité linguistique (au niveau des tokens, des phrases, des documents et de la temporalité) afin d'examiner les limites des modèles de langue lorsqu'ils sont entraînés sur des données provenant d'une seule source.

La deuxième contribution de cette thèse aborde les défis identifiés dans les modèles à source unique en explorant les *modèles de de langue médicaux à sources multiples*. Cette approche exploite des sources d'informations supplémentaires, notamment des bases de connaissances externes, des graphes de connaissances et des modèles de langue à grande échelle, afin de compléter les modèles à source unique et d'atténuer les limites spécifiques au domaine.

La troisième contribution est une étude réflexive sur l'écart de compétence entre les modèles de langue et les experts humains. Cette analyse prend en compte deux dimensions: (a) les questions centrées sur les modèles, en se concentrant sur les caractéristiques des modèles de langue dans le domaine médical, leurs choix de conception et les compromis qui surviennent lors du développement et du déploiement des modèles ; et (b) les questions centrées sur les données, en examinant les différences dans la compréhension des concepts, l'alignement des connaissances et la perception de la difficulté des tâches entre les modèles de langue et les experts humains.

Dans l'ensemble, cette thèse adopte une perspective critique sur les modèles de langue médicaux, fournissant des informations sur ce qui est nécessaire pour construire un « expert artificiel en santé » et soulignant les opportunités pour des recherches futures dans cette direction.

# Résumé long

Le langage médical constitue un sous-langage hautement spécialisé (Džuganová, 2019), qui se distingue par sa complexité inhérente, son utilisation intensive de terminologie spécifique au domaine et son degré élevé d'hétérogénéité entre les différents contextes cliniques et scientifiques. Contrairement au langage naturel général, le langage médical est souvent non structuré et fragmenté, contenant fréquemment des abréviations, du jargon et des références implicites qui nécessitent une expertise approfondie du domaine pour être interprétés avec précision. Cette expertise est généralement limitée aux professionnels de la santé, en particulier aux cliniciens et aux chercheurs médicaux, qui possèdent les connaissances de base nécessaires pour comprendre et contextualiser les informations. Par conséquent, le défi que représente le traitement du langage médical va au-delà de la compréhension conventionnelle du langage naturel, nécessitant des modèles capables de saisir efficacement les nuances et les spécificités propres au domaine médical.

Cette lacune dans la compréhension, due au manque d'expertise, se reflète également dans les performances des modèles d'apprentissage profond : les modèles pré-entraînés sur des corpus de domaine général (tels que des articles d'actualité, des textes sur internet ou des contenus encyclopédiques) ont souvent du mal à généraliser efficacement lorsqu'ils sont appliqués à des textes médicaux. Leur incapacité à s'adapter provient de différences fondamentales dans le vocabulaire, la syntaxe et la sémantique, ainsi que de la complexité contextuelle des récits médicaux.

Dans le chapitre 2 de la thèse, nous discutons donc en détail de la nature des données médicales et des modèles médicaux. Nous commençons par donner un aperçu du traitement automatique du langage médical, en résumant brièvement les pratiques dans le domaine du TAL. Nous passons ensuite à la description du traitement automatique du langage médical et discutons des lacunes des pratiques générales du TAL lorsqu'elles sont appliquées au domaine médical. Nous poursuivons avec les différentes origines des données médicales et, enfin, nous élaborons la nature des données médicales qui nous intéresse afin d'étudier les problèmes des modèles médicaux.

Nous nous concentrons sur les deux principaux facteurs qui expliquent pourquoi les modèles médicaux ont du mal à traiter les données médicales. Nous examinons la nature des ensembles de données disponibles dans le domaine médical et les techniques de modélisation utilisées. Les jeux de données médicales sont souvent plus petits, plus fragmentés et annotés de manière plus hétérogène que leurs équivalents dans le domaine général, ce qui entraîne des défis tels que la rareté des données, l'ambiguité des labels et les changements de domaine (Ogren et al., 2006; Ferraro et al., 2013; Bejan et al., 2012; Uzuner et al., 2011; Wicentowski and Sydes, 2008).

Nous étudions de manière exhaustive les défis liés à la nature des données médicales à l'aide d'une approche multifacette, en distinguant les données médicales en trois sources principales : les dossiers cliniques (y compris les dossiers médicaux électroniques, les notes cliniques et les résumés de sortie), la littérature scientifique (telle que

les articles biomédicaux et les revues systématiques) et les données médicales liées aux réseaux sociaux où les patients et les soignants partagent leurs expériences et des informations relatives à la santé. Chaque source présente des styles linguistiques, des degrés de structure et des contenus d'information distincts. Les textes cliniques sont généralement très techniques et semi-structurés, la littérature scientifique est formelle et fondée sur des preuves, tandis que le langage des réseaux sociaux est informel, bruyant et comprend souvent des expressions familières.

Ces différences nécessitent des techniques de traitement automatique des langues adaptées afin d'extraire, de représenter et d'interpréter avec précision les informations de chaque domaine. À leur tour, les choix de modélisation sont limités par ces caractéristiques des données, créant ainsi une boucle de rétroaction dans laquelle les limites des ensembles de données et les stratégies de modélisation influencent mutuellement leur efficacité respective.

Nous tirons parti de ces différences pour formuler notre première question de recherche (RQ 1; §3.1), qui porte sur les modèles médicaux à source unique n'utilisant qu'une seule des sources de données médicales mentionnées ci-dessus pour chaque tâche. Nous réalisons une analyse détaillée afin d'identifier les facteurs influençant les modèles médicaux lorsqu'ils sont entraînés avec une seule source de données. Nous étudions les niveaux linguistiques structurels et temporels des données médicales, y compris au niveau des tokens et des phrases, puis les aspects documentaires et temporels des données textuelles médicales (RQ 1.1-1.4; §3.2-3.5). Le chapitre 3 de la thèse présente cette exploration sous la forme d'une série de huit études de cas couvrant les niveaux linguistiques structurels et temporels des données médicales.

Au chapitre 3, nous présentons les principaux cas d'échec identifiés pour les modèles médicaux aux différents niveaux de complexité linguistique. Par exemple, les données médico-sociales utilisent souvent des expressions informelles ou un vocabulaire non standard, ce qui contraste fortement avec la terminologie médicale normalisée. Les données cliniques introduisent une complexité supplémentaire en raison de l'utilisation fréquente de raccourcis, d'abréviations, de détails contextuels riches, ainsi que de la variabilité des styles d'écriture. Les données médicales provenant de sources scientifiques, en revanche, se caractérisent par leur profondeur technique et une tendance à des styles d'écriture longs et à structure mixte.

Dans l'ensemble, une conclusion commune à toutes les études de cas est que les sources de données limitent à la fois la capacité de représentation et de généralisation des modèles médicaux. Dans l'ensemble, cette partie de la thèse met en évidence la manière dont la diversité du langage médical constitue un obstacle important pour les modèles formés à partir d'un seul type de source, soulignant l'importance de développer des approches robustes face à une telle variation.

La première série d'études de cas nous a permis de montrer que la richesse et la fiabilité des informations varient considérablement d'une source de données médicales à l'autre. Si la littérature scientifique fournit souvent des connaissances validées et éval-

uées par des pairs, les notes cliniques offrent des informations riches et spécifiques aux patients, essentielles pour la médecine personnalisée, et les données issues des réseaux sociaux permettent de saisir les expériences réelles des patients et les nouvelles tendances en matière de santé.

Cependant, la validité et la pertinence clinique des informations extraites des réseaux sociaux sont souvent incertaines, ce qui souligne la nécessité d'intégrer plusieurs sources de données afin d'améliorer la robustesse et la fiabilité. Cela incite à envisager le recours à des approches de fusion de données provenant de plusieurs sources. Le chapitre 4 explore donc cette piste dans le prolongement des fondements précédents, en formalisant la notion d'apprentissage multi-sources.

Nous examinons la deuxième question de recherche (RQ 2.1–2.3; §4.2-4.4), qui vise à déterminer si les modèles médicaux à sources multiples sont plus efficaces que leurs homologues à source unique, tout en mettant en évidence leur divergence par rapport aux approches à source unique. Contrairement aux modèles à source unique qui s'appuient sur un seul type de données, les modèles médicaux à sources multiples sont conçus pour exploiter plusieurs sources d'informations, dans le but d'atténuer les difficultés existantes. Au chapitre 4, nous examinons les trois orientations suivantes :

- Utilisation d'une base de connaissances externe ou d'un graphe de connaissances, qui offre des informations structurées et sélectionnées pouvant pouvant aider le modèle à l'inférence, garantissant ainsi que les prédictions sont fondées sur des concepts médicaux vérifiés plutôt que sur des associations erronées.

- Utilisation de la multimodalité, où les informations textuelles peuvent être enrichies par d'autres signaux tels que des images, des dossiers cliniques structurés ou des métadonnées associées aux les patients.

- Utilisation d'un grand modèle de langue (LLM), en tant qu'assistant formé à partir d'une abondance de connaissances préalables afin de surmonter les caractéristiques contextuelles nuancées des données médicales.

Grâce à une approche d'apprentissage multi-sources, nous visons à remédier aux limites identifiées des modèles de langage à source unique, permettant ainsi un traitement plus fiable des textes médicaux hétérogènes et soutenant des modèles de langage médicaux plus sensibles au contexte.

En outre, le chapitre 4 met l'accent sur l'évaluation de ces systèmes à sources multiples dans des contextes cliniques réalistes afin d'évaluer leur utilité pratique et leur impact sur la pratique médicale. Grâce à cette évaluation, nous souhaitons comprendre non seulement les améliorations en termes de précision prédictive par rapport aux modèles de langage médicaux à source unique, mais également la manière dont ces systèmes influencent les flux de travail cliniques, la confiance dans les décisions et, en fin de compte, les résultats pour les patients.

Nos résultats suggèrent que l'apprentissage à sources multiples offre une voie convaincante vers une robustesse et une compréhension contextuelle améliorées en traitant

les problèmes liés aux données. En tirant parti de données hétérogènes et d'une structure externe, ils commencent à combler le fossé entre la prédiction computationnelle et les connaissances de niveau expert. Cependant, la complexité de l'intégration de multiples sources de connaissances soulève également plusieurs défis pour les modèles de langage, notamment des risques tels qu'une confiance incohérente, une dérive des connaissances et une perte d'interprétabilité.

Dans la continuité de la fin du chapitre précédent, le chapitre 5 passe de la modélisation des phénomènes médicaux à une réflexion sur les défis plus larges qui restent à relever pour les modèles de langage médical par rapport aux processus de raisonnement des experts médicaux. Alors que les chapitres précédents examinaient l'apprentissage à source unique et à sources multiples comme des stratégies concrètes pour construire des systèmes plus robustes, nous adoptons ici une perspective plus critique concernant les limites qui persistent malgré ces progrès.

Une préoccupation centrale est l'incertitude/la confiance excessive des modèles de langage, qui reste inhérente aux paradigmes de modélisation actuels et pose des défis pour un déploiement clinique fiable, entraînant un déficit de compétences par rapport aux experts médicaux. La confiance excessive peut résulter d'hypothèses de modélisation, d'un surajustement à des ensembles de données spécifiques ou de biais d'annotation, pouvant conduire à des recommandations cliniques erronées aux conséquences graves.

Cette situation est encore aggravée par le compromis avec l'interprétabilité, car les modèles de langage qui obtiennent des performances prédictives plus élevées le font souvent au détriment de la transparence, ce qui rend leurs prédictions difficiles à aligner sur les processus de raisonnement clinique. Enfin, au-delà de ces défis, nous nous penchons sur une interaction plus complexe entre l'incertitude et la complexité des tâches inhérentes à la modélisation des données, que les approches actuelles n'ont pas encore pleinement prise en compte.

Plus concrètement, nous réfléchissons aux principaux défis qui continuent de limiter les progrès et qui conduisent à l'écart observé entre les modèles de langage médicaux et les experts médicaux (humains).

- Premièrement, nous réfléchissons aux questions liées aux modèles de langage médicaux, qui concernent l'incertitude inhérente aux données médicales et la nature opaque des approches fondées sur les données.

- Ensuite, nous examinons les problèmes liés aux données, en particulier les problèmes de variation et d'alignement liés à l'étiquetage humain et à la compréhension conceptuelle, qui, même parmi les cliniciens experts, introduisent des incohérences, compliquant le développement et l'évaluation des modèles.

Les études de cas présentées dans ce chapitre traitent individuellement de ces défis et offrent un aperçu de leurs causes sous-jacentes afin de mettre en perspective l'importance d'une conception prudente des modèles de langage médicaux. Il est essentiel de reconnaître et d'ajuster la confiance excessive grâce à une quantification

rigoureuse de l'incertitude, à une interprétabilité transparente des modèles de langage et à des pratiques d'annotation minutieuses pour déployer une IA fiable dans le domaine des soins de santé.

Cette réflexion souligne la nécessité d'un changement de paradigme vers la création de systèmes de TAL médicaux qui soient non seulement précis, mais aussi fiables, interprétables et adaptés aux réalités complexes de l'expertise clinique et de la prise de décision. En outre, en examinant ces questions, ce chapitre replace les contributions techniques de la thèse dans un contexte plus large et pose les bases de la discussion finale, où leurs implications pour la recherche future et l'intégration médicale sont synthétisées.

Nous concluons notre étude par le chapitre 6, qui rassemble les principales conclusions identifiées tout au long de la thèse, avec l'étude des modèles de langage à source unique, l'exploration des stratégies d'apprentissage à sources multiples et, enfin, la réflexion critique sur l'écart de compétence entre les modèles de langage médicaux et les experts médicaux humains. Il synthétise ces contributions en une perspective cohérente sur l'état de la compréhension du langage médical, en soulignant à la fois les progrès méthodologiques réalisés et les limites qui subsistent.

La thèse partait d'une question simple mais fondamentale : « Pourquoi les modèles de langage médicaux ont-ils encore du mal à égaler les capacités de raisonnement des experts médicaux ? » L'exploration qui a suivi a révélé qu'il ne s'agissait pas simplement d'une question d'entraînement sur davantage de données ou de construction de modèles plus grands. Il s'agit d'un défi multiforme impliquant la complexité linguistique, la nature temporelle et les biais systémiques tant dans les modèles que dans les données. Ce faisant, elle identifie les possibilités d'intégrer plus directement l'expertise médicale dans la conception des modèles, afin de développer des modèles de raisonnement médical plus robustes.

Concrètement, cette recherche contribue au développement de modèles de langage médicaux plus sûrs et plus robustes. Les connaissances acquises sur l'intégration de connaissances externes, de données multimodales et de l'utilisation assistée par LLM peuvent éclairer la conception d'architectures hybrides qui offrent à la fois des performances et des résultats interprétables. De plus, l'accent mis sur les diagnostics centrés sur les données et les modèles fournit une base pour le raffinement itératif des modèles de langage, l'évaluation comparative spécifique au domaine et les stratégies de déploiement impliquant l'intervention humaine.

Bien que cette thèse explore de manière exhaustive la position relative des modèles de langage médicaux et des experts médicaux, il convient de reconnaître certaines limites. Premièrement, les études de cas reposaient sur des conceptions expérimentales relativement simples et excluaient en grande partie les grands modèles linguistiques, en raison à la fois du calendrier des travaux et de la volonté de conserver une perspective critique sur un paysage de recherche en rapide évolution.

La thèse ne couvre pas entièrement les développements les plus récents en matière

de grands modèles de langages médicaux, notamment les modèles de base et de raisonnement médicaux. Deuxièmement, les études s'appuient sur des ensembles de données médicales accessibles au public et souvent de petite taille, en anglais, français et espagnol, qui ne reflètent peut-être pas entièrement la diversité et la complexité du langage clinique réel.

Troisièmement, bien que plusieurs dimensions du décalage entre les modèles et les experts aient été étudiées, les comparaisons quantitatives avec les annotations humaines sont restées limitées en raison de l'accès restreint aux experts du domaine. Enfin, l'intégration de connaissances externes et de modalités supplémentaires a été évaluée dans des conditions expérimentales contrôlées, alors que le déploiement dans le monde réel impliquerait d'autres défis, notamment la synchronisation des données, les contraintes de confidentialité, l'évolutivité du système et le changement de distribution.

Dans l'ensemble, cette thèse préconise de dépasser les processus d'évaluation traditionnels pour s'orienter vers des cadres qui tiennent compte de l'incertitude, de l'interprétabilité, de la modularité et de l'alignement des domaines. Si les modèles de langage ont un immense potentiel pour améliorer les flux de travail médicaux, leur fiabilité doit être acquise non seulement par leur précision, mais aussi par leur transparence et leur compatibilité avec les valeurs de la pratique médicale.

À l'avenir, les recherches sur les modèles de langage médicaux devront relever plusieurs défis persistants qui entravent leur déploiement sûr et efficace dans les milieux cliniques. Il s'agit notamment des hallucinations, du désalignement entre les modèles et les experts, et de l'adaptation inadéquate au domaine, qui sont encore aggravés par des cadres d'évaluation défaillants pouvant encourager le recours à des corrélations fallacieuses, la génération d'informations cliniques incorrectes et l'absence de mécanismes fiables pour exprimer l'incertitude dans des scénarios à haut risque.

De plus, des connaissances médicales structurées limitées, une interprétabilité insuffisante et une vulnérabilité à la dérive du domaine peuvent donner lieu à des modèles qui semblent fluides mais qui restent peu fiables. Pour remédier à ces problèmes, il faudra s'orienter vers des systèmes qui soient non seulement précis, mais aussi fiables, transparents et robustes face à la complexité du raisonnement clinique, tout en développant des cadres d'évaluation rigoureux capables de surveiller le comportement des modèles et de tenir compte de la subjectivité des annotations humaines, notamment grâce à des approches telles que les LLM en tant qu'évaluateurs et des outils tels que LM-Polygraph.

Ensemble, ces réflexions et propositions marquent la conclusion des travaux présentés dans cette thèse, tout en ouvrant la voie à des progrès continus à l'intersection du traitement automatique des langues et de la pratique médicale.

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

**AI** Artificial Intelligence. xxvi, xxxiii, 5, 14, 139, 158–161, 172, 182, 221

**ANOVA** Analysis of variance. xxvi, 167

**AUPRC** Area Under the Precision-Recall Curve. xxvi, 103–106

**AUROC** Area Under the Receiver Operating Characteristic Curve. xxvi, 103, 105, 106

**BARTHEZ** French Biomedical Language Model based on BART. xxvi

**BEA** Building Educational Applications. xxvi, 69, 148

**BERT** Bidirectional Encoder Representations from Transformers. xxvi, 18–20, 49, 60, 69, 81, 122, 141, 174, 175, 184, 207, 214

**BLEU** Bilingual Evaluation Understudy. xxvi, 55, 56, 123

**BLEURT** Bilingual Evaluation Understudy with Representations from Transformers. xxvi, 55, 56, 123, 127

**BLUE** Biomedical Language Understanding Evaluation. xxvi, 184

**BLURB** Biomedical Language Understanding Benchmark. xxvi, 184

**BNN** Bayesian Neural Network. xxvi, 163, 165

**BS** Brier Score. xxvi, 163, 164, 166

**CE** Cross-Entropy. xxvi

**CIR** Class Imbalance Ratio. xxvi

**CLIN** Clinical. xxvi, 9, 10, 24, 37, 52, 112, 137

**CLS** Classification Token [CLS]. xxvi, 81, 82, 190, 192

**CME2** Contextual Medication Event Extraction. xxvi

**CMED** Contextualized Medication Event Dataset. xxvi, 48, 62, 68, 148

**CN** Clinical Notes. xxvi, 139

**CNN** Convolutional Neural Network. xxvi, 40, 138

**CONCAT** Concatenation-Based Baseline. xxvi, 138

**CP** Conformal Prediction. xxvi, 204, 208, 209, 212, 213

**CW** Class Weights. xxvi, 64

**DEV** Development Set. xxvi

**DIS** Disease Entity. xxvi

**DL** Deep Learning. xxvi, 14

**DNN** Deep Neural Network. xxvi, 163

**DP** DropOut. xxvi

**DW** DeepWalk. xxvi, 132, 133

**E2E** End-to-End. xxvi

**ECE** Expected Calibration Error. xxvi, 163, 164, 166–168

**EFL** English as a Foreign Language. xxvi, 67

**EHR** Electronic Health Record. xxvi, 47, 61, 86

**EHRs** Electronic Health Records. xxvi, 97, 161

**ELMo** Embeddings from Language Models. xxvi, 17, 18, 20

**EMP** English for Medical Purposes. xxvi, 67

**EMR** Electronic Medical Record. xxvi, 68, 148

**ESMO** European Society for Medical Oncology. xxvi, 186

**FLAIR** Fast, Lightweight and Accurate Textual Entailment Recognizer. xxvi, 43, 115

**FPR** False Positive Rate. xxvi, 103

**FT** Fine-Tuning. xxvi, 56

**GloVe** Global Vectors for Word Representation. xxvi, 17, 19, 20, 48, 49, 62, 63, 65

**GPT** Generative Pretrained Transformer. xxvi, 18, 19, 21, 91, 150, 184

**GPTQ** Quantized Generative Pretrained Transformer. xxvi, 121

**GRU** Gated Recurrent Unit. xxvi, 18, 21

**GS** GraphSage. xxvi, 132, 133

**HoI** Hallmark of Immunotherapy. xxvi, 187, 189, 190, 192, 193, 195, 196

**ICD** International Classification of Diseases. xxvi, 113

**ICU** Intensive Care Unit. xxvi, 29

**IG** Integrated Gradients. xxvi, 174, 176, 177, 181, 182

**IID** Independent and Identically Distributed. xxvi, 38

**IR** Information Retrieval. xxvi

**ISTS** Irregularly Sampled Time Series. xxvi, 97–99, 102, 103, 107, 139

**KB** Knowledge Base. xxvi, 123

**KG** Knowledge Graph. xxvi

**kNN** k-Nearest Neighbor. xxvi, 205

**LLM** Large Language Model. xii, xxvi, 46, 52, 58, 71, 89, 91, 92, 111, 112, 145–148, 150–152, 154, 170, 184, 190–197, 207, 221–224

**LM** Language Model. xxvi, 184, 190, 195, 223, 224

**LM** Language Model. xxvi, 43, 53

**LOC** Location Entity. xxvi

**LoRA** Low-Rank Adaptation. xxvi, 141

**LSTM** Long Short-Term Memory. xxvi, 17, 18, 21, 49, 63, 64, 98–101, 105, 107, 138–140

**LSTM** Long Short-Term Memory. xxvi

**LTM** Long-Term Memory. xxvi, 99, 100, 105, 106

**MCD** Monte Carlo Dropout. xxvi, 163

**MED** Medication Entity. xxvi

**ML** Machine Learning. xxvi, 6

**MLM** Masked Language Modeling. xxvi, 190

**MLU** Medical Language Understanding. xxvi, 14

**MNAR** Missing Not At Random. xxvi, 137

**mSLAN** Multi-modal Switch LSTM Aggregation Network. xxvi, 139, 141–143

**NER** Named Entity Recognition. xxvi, 40, 41, 43–45, 47, 60, 81, 118

**NLI** Natural Language Inference. xxvi, 173, 178, 180, 181, 206, 207

**NLL** Negative Log-Likelihood. xxvi, 164, 166

**NLP** Natural Language Processing. xxvi, 6, 9, 13, 14, 16, 18–21, 23, 24, 26–28, 30, 31, 34, 38, 40, 41, 53, 60, 71, 79, 85, 92, 113, 131, 135, 136, 145, 146, 159, 161, 170, 172, 173, 184–186, 200, 201, 214, 220, 223

**NLU** Natural Language Understanding. xxvi, 14

**OLMo** Open Language Model. xxvi, 207

**OLS** Ordinary Least Squares. xxvi, 69, 150

**OOD** Out-of-Distribution. xxvi, 160

**OOV** Out of Vocabulary. xxvi, 41

**ORG** Organization Entity. xxvi

**PaLM** Pathways Language Model. xxvi, 21

**PER** Person Entity. xxvi

**PLM** Pretrained Language Model. xxvi, 69, 81, 82, 88–92, 112, 161, 163, 165, 190, 192–194, 196, 212, 214, 224

**POS** Part of Speech. xxvi, 62

**pRAGe** Pipeline for Retrieval Augmented Generation and Evaluation. xxvi, 121–127

**PROC** Procedure Entity. xxvi

**Q-LoRA** Quantized Low-Rank Adaptation. xxvi, 54, 121

**RAG** Retrieval-Augmented Generation. xxvi, 85, 123, 124

**RNN** Recurrent Neural Network. xxvi, 17, 21, 49, 61, 63, 64

**ROC** Receiver Operating Characteristic. xxvi, 103

**ROUGE** Recall-Oriented Understudy for Gisting Evaluation. xxvi, 55, 56, 123

**RSTS** Regularly Sampled Time Series. xxvi, 26, 27, 98, 102, 103

**SCE** Static Calibration Error. xxvi, 163, 164, 166, 167

**SCI** Scientific. xxvi, 9, 10, 24, 37, 83, 112, 134, 185

**SDF** Spanish Disease Finder. xxvi, 45

**SEP** Separator Token. xxvi

**SHAP** SHapley Additive exPlanations. xxvi, 167, 169, 174, 176, 177, 181, 182

**SLAN** Switch LSTM Aggregation Network. xxvi, 98, 99, 101, 103–107, 139, 141–143

**SLM** Small Language Model. xxvi, 88

**SOC** Social. xxvi, 9, 10, 24, 37, 52, 112, 114, 118

**SOTA** State-of-the-art. xxvi, 123

**SST2** Stanford Sentiment Treebank. xxvi, 173, 174

**ST** Sentence Truncation. xxvi, 65

**STM** Short-Term Memory. xxvi, 99, 100

**SYM** Symptom Entity. xxvi

**T2D** Type 2 Diabetes. xxvi, 68, 148

**TAL** traitement automatique des langues. viii, xxvi

**TF-IDF** Term Frequency-Inverse Document Frequency. xxvi, 81, 82, 133

**TNM** Tumor, Node, Metastasis Cancer Staging. xxvi, 86, 89, 92

**TPR** True Positive Rate. xxvi, 103

**TS** Time Series. xxvi, 139

**tsSLAN** Time Series SLAN. xxvi, 141–143

**txtSLAN** Text-SLAN. xxvi, 141–143

**ULMFiT** Universal Language Model Fine-tuning. xxvi, 18

**UMLS** Unified Medical Language System. xxvi, 31, 32, 113, 114, 123

**UTDE** Unified Temporal Discretization-based Embedding. xxvi, 141, 143

**VI** Variational Inference. xxvi, 163

**WE** Word Embedding. xxvi, 64

**WMN** Wikineural Multilingual NER. xxvi

# Use of AI Tools

I hereby declare that Artificial Intelligence (AI) tools, such as OpenAI's ChatGPT and Claude, were used to assist in the preparation of this thesis in two primary ways: writing refinement and technical formatting support.

**Writing and Refinement.** AI was employed to improve clarity, grammar, and the formal tone of selected paragraphs. It also provided suggestions for restructuring sentences and enhancing readability. We provide an instance below with the prompt in gray and generated output in blue.

`Refine this` : "*While, deep-learning based methods such as hybrid model using RNNs and residual network (Rumeng et al., 2017); multitask learning (Bhatia et al., 2018) suffer mostly ambiguity caused by writing styles (such as misspellings, abbreviations, inconsistent tense usage) in EHRs.*"

*Meanwhile, deep learning approaches, including hybrid RNN-residual network architectures (Rumeng et al., 2017) and multitask learning frameworks (Bhatia et al., 2018), are more robust but still face challenges. These include ambiguity arising from diverse writing styles in electronic health records (EHRs), such as misspellings, abbreviations, and inconsistent tense usage.*

**Technical Formatting Support.** AI was used to occasionally generate illustration plots from LaTeX tables or change the structure of tables or plots. For example : To obtain a visual illustration of term variation distance:

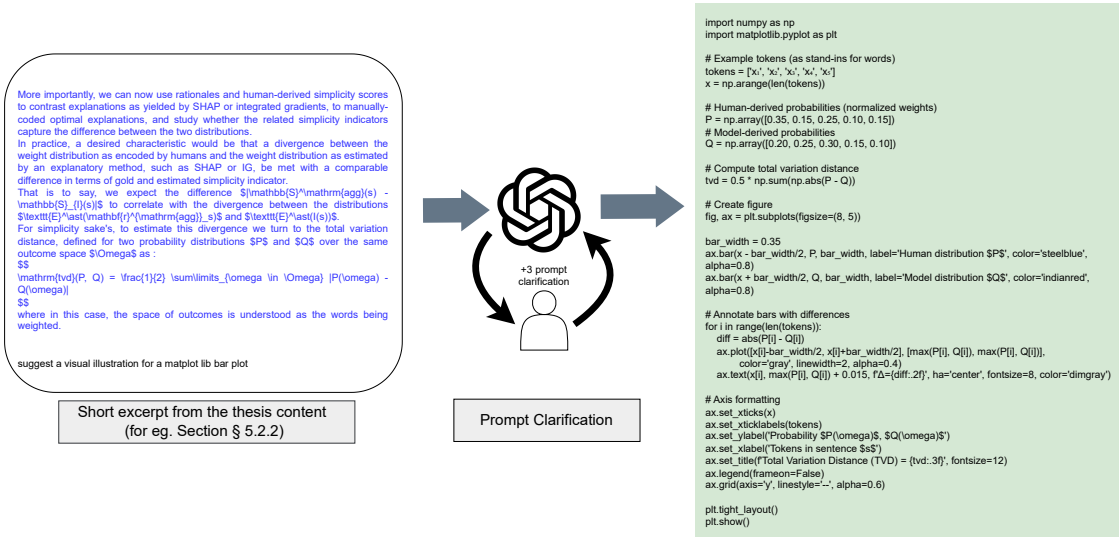



Figure 1: Illustration of use of GPT to obtain visualization of the thesis. The obtained figure can be seen in §5.2.2, Figure 5.10.

All content including AI enhanced text, figures, tables, and visual elements reviewed, revised and finalized by the author to ensure accuracy, factuality, originality, and consistency with the academic standards of this work.
This declaration is made in the interest of transparency and academic integrity. The author remains fully responsible for the intellectual content, analysis, and conclusions presented in this thesis.

# Part I

# Where Words Meet Medicine

# Chapter 1

# Introduction



This chapter sets the stage for the thesis by framing the research within its broader context and outlining the central questions that drive the investigation. It then summarizes the key scientific and community contributions, highlighting both the methodological advances and their practical relevance. Finally, the chapter provides a structured overview of the thesis, clarifying how each subsequent chapter builds upon these questions and contributions to form a coherent narrative.

## 1.1 Thesis Context

Medical language constitutes a highly specialized sublanguage (Džuganová, 2019), distinguished by its inherent complexity, extensive use of domain-specific terminology, and a high degree of heterogeneity across different clinical and scientific contexts. Unlike general natural language, medical language is often unstructured and fragmented, frequently containing abbreviations, jargon, and implicit references that require substantial domain expertise to interpret accurately. This expertise is typically limited to healthcare professionals, particularly clinicians and medical researchers, who possess the necessary background knowledge to understand and contextualize the information. Consequently, the challenge of processing medical language extends beyond conventional natural language understanding, requiring models that can effectively capture the nuances and specificity unique to the medical domain.

This gap in understanding is reflected in the performance of deep learning models: models pretrained on general-domain corpora—such as news articles, web text, or encyclopedic content—often struggle to generalize effectively when applied to medical texts. Their inability to adapt stems from fundamental differences in vocabulary, syntax, and semantics, as well as the contextual complexity of medical narratives. Two principal factors contribute to this issue: the nature of the datasets available in the medical domain and the modeling techniques employed. Medical datasets are often smaller, more fragmented, and more heterogeneously annotated compared to their general-domain counterparts, leading to challenges such as in data sparsity, label noise, and domain shifts (Ogren et al., 2006; Ferraro et al., 2013; Bejan et al., 2012; Uzuner et al., 2011; Wicentowski and Sydes, 2008). In turn, modeling choices are constrained by these data characteristics, creating a feedback loop where dataset limitations and modeling strategies mutually influence each other's effectiveness.

To address these challenges comprehensively, this thesis adopts a multi-faceted approach by categorizing medical data into three primary sources: clinical records (including electronic health records, clinical notes, and discharge summaries), scientific literature (such as biomedical articles and systematic reviews), and social media content where patients and caregivers share experiences and health-related information. Each source presents distinct linguistic styles, degrees of structure, and information content. Clinical texts are typically highly technical and semi-structured, scientific literature is formal and evidence-based, while social media language is informal, noisy, and often includes layman expressions. These differences require tailored natural language processing techniques to accurately extract, represent, and interpret information from each domain.

Moreover, the information richness and reliability vary considerably among these sources. While scientific literature often provides validated and peer-reviewed knowledge, clinical notes offer rich patient-specific insights that are critical for personalized medicine, and social media data can capture real-world patient experiences and emerging health trends. However, the validity and clinical relevance of information extracted from social media are often uncertain, highlighting the need for integrating multiple data sources to enhance robustness and reliability. By leveraging multi-source data

fusion, this thesis explores the development of clinical decision support systems that combine individualized patient data with aggregated evidence from broader populations, thus enabling more holistic and context-aware monitoring and decision-making.

Furthermore, this work emphasizes evaluating such multi-source systems in realistic clinical settings to assess their practical utility and impact on medical practice. Through this evaluation, we aim to understand not only the improvements in predictive accuracy but also how these systems influence clinical workflows, decision confidence, and ultimately patient outcomes.

Finally, reflecting on the broader implications of our findings, this thesis critically examines the phenomenon of overconfidence in clinical decision support systems. Overconfidence can arise due to modeling assumptions, overfitting to specific datasets, or annotation biases, potentially leading to erroneous clinical recommendations with serious consequences. Recognizing and mitigating overconfidence through rigorous uncertainty quantification, transparent model interpretability, and careful annotation practices is crucial for the deployment of trustworthy AI in healthcare. This reflection underscores the need for a paradigm shift towards building medical NLP systems that are not only accurate but also reliable, interpretable, and aligned with the complex realities of clinical expertise and decision-making.

## 1.2 Research Questions

In this section, we describe the different research questions that we attend to in the context of building medical models that can be helpful to clinical decision support system. The research questions are framed from the perspective of medical language understanding in order to investigate the nuances that impact medical models when dealing with medical text data.

RQ 1 (§ 3.1)

What are the challenges faced by Single-Sourced Medical Models?

The application of Natural Language Processing (NLP) in medical domain has uplifted tremendously the automation of complicated tasks needed that clinicians come across on daily basis. This can be attributed to deep learning where access to sufficient or enough data can be used to create a model that outperforms traditional Machine Learning (ML) or NLP models. This approach was dependent mostly on one source of data and therefore we in the thesis referred to such models as single sourced medical models. In order to investigate this research question, we explore and identify what are the different situations where medical models struggle when they are trained with one source of data. This involves the following list of investigation studies :

RQ 1.1 **What are the challenges medical models face on word-level?** (§ 3.2)

RQ 1.2 **What are the challenges medical models face on sentence-level?** (§ 3.3)

RQ 1.3 **What are the challenges medical models face on document-level?** (§ 3.4)

RQ 1.4 **What are the challenges medical models face on temporal-level?** (§ 3.5)

RQ 2 (§ 4.1)

Are Multi-Sourced Medical Models more effective compared to Single-Sourced Medical Models for medical language processing?

Building on the limitations identified in the preceding research question regarding single-sourced medical models, we extend our investigation to assess the extent to which integrating heterogeneous and complementary sources of information can mitigate these shortcomings. Specifically, we posit that a multi-sourced medical modeling approach has the potential to enhance robustness, improve generalizability across diverse clinical settings, and reduce the biases inherent to single-domain data. To evaluate this hypothesis, we design and conduct a series of case studies that systematically examine the practical benefits and challenges of multi-sourced modeling in representative medical contexts. The case studies are presented as follows:

RQ 2.1 **Can the use of external knowledge base help to mitigate challenges faced by single-sourced models?** (§ 4.2)

RQ 2.2 **Can the use of additional modality help to mitigate challenges faced by single-sourced models?** (§ 4.3)

RQ 2.3 **Can the use of large language models (LLMs) as an assistant help to mitigate challenges faced by single-sourced models?** (§ 4.4)

RQ 3 (§ 5.1)

Can we explain the proficiency gap between Medical Models and Medical Experts?

Although recent advances in medical AI have demonstrated considerable promise, models deployed in real-world clinical settings remain far from attaining the proficiency of human medical experts. With this research question, we reflect on the central challenges that continue to limit progress and can be attributed to the proficiency gap between medical models and medical experts. First, we reflect onto the model centric related issues of medical models that is concerned with the inherent uncertainty in medical data and blackbox nature of data driven approaches. And second, we consider the data related issues in particular, the variation and alignment problems around human labeling and conceptual understanding, which even among expert clinicians introduces inconsistencies, complicating model development and evaluation. The studies presented in this chapter cover these challenges individually, offering insights into their underlying causes to put in perspective the importance of cautious design of medical models.

RQ 3.1 **What are the model-related issues contributing to the gap between models and experts?** (§ 5.2)

RQ 3.2 **What are the data-related issues contributing to the gap between models and experts?** (§ 5.3)

## 1.3 Scientific Contributions

The scientific contributions of this thesis are consolidated through a series of peer-reviewed publications. Each of these works addresses a distinct aspect of the research questions explored, while collectively building towards the overarching objectives of the dissertation. The following papers represent the primary outputs, where I have played a central role in conceptualization, methodology, and analysis. Together, they demonstrate the progression of ideas, the methodological innovations proposed, and the application of natural language processing techniques to key problems in the medical domain.

1. **Sinha, A.**, Holgado, C. G., Clausel, M., & Constant, M. (2022). IAI@ SocialDisNER: Catch me if you can! Capturing complex disease mentions in tweets. In Mining for Health Applications, Workshop & Shared Task (# SMM4H 2022) (p. 85).

2. **Sinha, A.**, Vishwakarma, A., Clausel, M., & Constant, M. (2023). CME² Net: Contextual Medical Event Extraction Network for clinical notes. In CEUR Workshop Proceedings (Vol. 3416, pp. 23-29).

3. **Sinha, A.**, Bigeard, S., Clausel, M., & Constant, M. (2023). What shall we read: the article or the citations?-A case study on scientific language understanding. In Actes de CORIA-TALN 2023. Actes de l'atelier "Analyse et Recherche de Textes Scientifiques" (ARTS)@TALN 2023, pages 80–85, Paris, France. ATALA.

4. Agarwal, R., **Sinha, A.**, Prasad, D. K., Clausel, M., Horsch, A., Constant, M., & Coubez, X. (2023). Modelling Irregularly Sampled Time Series Without Imputation. arXiv preprint arXiv:2309.08698.

5. **Sinha, A.**, Holgado, C. G., Clausel, M., Constant, M., & Coubez, X. (2023). Can LLMs be used to understand clinical notes better?. Actes des 5èmes journées du Groupement de Recherche CNRS "Linguistique Informatique, Formelle et de Terrain", 94.

6. **Sinha, A.**, Mickus, T., Clausel, M., Constant, M., & Coubez, X. (2024, August). Domain-specific or Uncertainty-aware models: Does it really make a difference for biomedical text classification?. In Proceedings of the 23rd Workshop on Biomedical Natural Language Processing, pages 202–211, Bangkok, Thailand. Association for Computational Linguistics.

7. Buhnila, I., **Sinha, A**., & Constant, M. (2024, August). Retrieve, Generate, Evaluate: A Case Study for Medical Paraphrases Generation with Small Language Models. In Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024), pages 189–203, Bangkok, Thailand. Association for Computational Linguistics.

8. **Sinha, A.**, & Buhnila, I. (2025, June). ATILF at NTCIR-18 RadNLP 2024 Shared Task: With less radiology reports, comes less performance. In NTCIR-18: Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies (pp. 354-358).

9. **Sinha, A.** (2025). IAI at #SMM4H-HeaRD 2025: Benchmarking PLMs for medical language understanding tasks. In Proceedings of the #SMM4H-HeaRD 2025: Joint 10th Social Media Mining for Health and Health Real-World Data Workshop and Shared Tasks, Copenhagen, Denmark.

10. Mickus, T., **Sinha, A.**, & Vázquez, R. (2025). Your model is overconfident, and other lies we tell ourselves. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5401–5417, Vienna, Austria. Association for Computational Linguistics.

11. **Sinha, A.**, Mickus, T., Clausel, M., Constant, M., & Coubez, X. (2025a). Simplicity isn't as simple as you think. In The 1st Workshop on Actionable Interpretability at Forty-Second International Conference on Machine Learning, Vancouver, Canada.

12. **Sinha, A.**, Popescu, B.-V., Coubez, X., Clausel, M., & Constant, M. (2025b). Immunofomo: Are language models missing what oncologists see? arXiv preprint arXiv: 2506.11478.

## 1.4 Community Contributions

In addition to the written contributions, this thesis also reflects my active involvement in the scientific community through the organization of workshops, symposia, and related events. These activities provided a platform for disseminating knowledge, fostering interdisciplinary dialogue, and facilitating collaboration among researchers and practitioners. The following list highlights the key scientific events I have contributed to organize, underscoring my commitment to advancing research exchange and community building within the field.

1. Buhnila, I., **Sinha, A.**, Song, H., Zanella, L., Klein, S., Laroche, J., Charuau, D., & Bigeard, S. (2025, June). MLP-LLM : Workshop on Medical Language Processing in the era of Large Language Models collocated with CORIA-TALN 2025, Marseille, France. URL : https://atilf-umr7118.github.io/MLPLLM2025/

2. **Sinha, A.**, Mickus, T., Vazquez, R., Buhnila, I., Agarwal, R., Schmidtova, P., Tiedmann, J., & Prasad, D. K. (2025, December). CHOMPS : Workshop on Confabulation, Hallucination, Over-generation in Multilingual and Precision Settings collocated with IJCNLP-AACL 2025, Mumbai, India. URL : https://chomps2025.github.io/

## 1.5 Thesis Outline

The thesis is structured as follows:

**Chapter 2: Background.** This chapter intends to provide the pre-requisites that are needed to follow the course of the topics that will be discussed in the following chapters. We provide an overview of medical language processing covering the brief summary of practices in the field of natural language processing (NLP). We then move on to describe medical language processing and discuss how the general NLP practices fall short when applied to medical domain. We discuss the different origins of medical data and divide them into three separate sources namely Social (SOC) media related, Clinical (CLIN) related and Scientific (SCI) publication related. Finally, we discuss the nature of medical data which is point of interest in order to investigate the problems in clinical models.

**Chapter 3: Single-sourced Medical models.** This chapter focuses on the use of single-source learning for medical models and examine its implications on medical downstream tasks. As a starting point, we identify the key challenges that arise when models trained on a single source are applied to diverse forms of medical text such as from [SOC], [CLIN] and [SCI] data sources. These challenges stem largely from the heterogeneous nature of medical data. For example, [SOC] medical data often employs informal expressions or non-standard vocabulary, which contrasts sharply with standardized medical terminology. [CLIN] data introduces additional complexity through the frequent use of shorthand, abbreviations, rich contextual detail, as well as variability in writing style. [SCI] source originated medical data, on the other hand, is characterized by its technical depth, and a tendency towards lengthy, mixed-structure writing styles. Taken together, these factors highlight how the diversity of medical language poses significant obstacles for models trained on a single type of source, underlining the importance of developing approaches that are robust to such variation.

**Chapter 4: Multi-sourced Medical models.** This chapter, in continuation of the previous chapter, firstly formalizes the notion of multi-source learning and highlights the divergence from single-source approaches. In contrast to single-sourced models that rely on one type of data, multi-sourced medical models are designed to draw upon more than one source of information, thereby aiming to mitigate existing challenges. We explore the following promising directions (i) the use of external knowledge base which offers structured, curated information that can supplement model training and inference, ensuring that predictions are grounded in verified medical concepts rather than spurious associations. (ii) the use of Multi-modality, where textual information can be enriched with other signals such as images, structured clinical records, or patient metadata. (iii) the use of large language model as an assistant trained with abundance of prior knowledge to overcome nuanced contextual characteristics of medical data. By integrating these diverse sources, multi-sourced learning aims to address the inherent limitations of single-source models, enabling more reliable handling of heterogeneous medical texts and supporting deeper, contextually aware medical models.

**Chapter 5: Medical Models vs. Medical Experts.** In this reflective chapter, we shift the focus from modeling medical phenomena to reflecting on the broader challenges that remain for medical language models when compared to the reasoning processes of medical experts. While the preceding chapters examined single-source and multi-source learning as concrete strategies for building better and more robust systems, here we adopt a more critical perspective on the limitations that persist despite these advances. A central concern is uncertainty, which remains inherent to the current modeling paradigms and raises challenges for dependable clinical deployment. This is further compounded by the trade-off with interpretability, as models that achieve higher predictive performance often do so at the expense of transparency, making their predictions difficult to align with clinical reasoning processes. Finally, beyond these challenges we look into a more intricate interplay between uncertainty and task complexity inherent to data modeling, which current approaches have yet to fully address. By examining these issues, this chapter situates the technical contributions

of the thesis within a broader context and lays the groundwork for the concluding discussion, where their implications for future research and medical integration are synthesized.

**Chapter 6: Conclusion and Future Work.** This chapter brings together the main findings identified across the course of the thesis, with the investigation of single-source models, the exploration of multi-source learning strategies, and finally the critical reflection on the proficiency gap between the medical models and medical human experts. It synthesizes these contributions into a coherent perspective on the state of medical language understanding, emphasizing both the methodological advances achieved and the limitations that remain. Beyond summarizing the key findings, the chapter also looks forward, outlining directions for future work that could lead to better medical language understanding via medical language models. In doing so, it identifies opportunities for integrating medical expertise more directly into model design, for developing more robust medical reasoning models. Together, these reflections and proposals mark the conclusion of the present work while opening pathways for continued progress at the intersection of natural language processing and medical practice.

# Chapter 2

# Background



In this chapter, we provide an overview of the foundational background necessary to understand the subsequent chapters and the studies presented in this thesis. We begin by introducing medical language processing, starting with a brief history of automatic language processing in Section (§2.1.1), followed by a discussion of the unique challenges that make medical language particularly difficult for NLP models. In Section (§2.2), we describe the various types and sources of medical data, highlighting their distinctive characteristics and the challenges they present for machine learning and deep learning approaches. Finally, in Section (§2.3), we explore how language models interact with medical data, discussing model limitations, domain adaptation issues, knowledge integration strategies, and direction towards multi-source model modeling.

## 2.1 Medical Language Processing

Medical language processing, often referred to as *clinical language processing*, is a specialized branch of NLP focused on extracting structured insights from medical text data. Unlike general natural language processing, medical NLP deals with domain-specific terminology, linguistic patterns, and the scarcity of annotated data[1] (Wang et al., 2018b). These challenges require the use of models that go beyond shallow text pattern matching by incorporating contextual and semantic understanding of medical knowledge through methods in artificial intelligence (AI) and Deep Learning (DL) (Hirschberg and Manning, 2015; Choi et al., 2016). It is important to note that, the terms Medical Language Understanding (MLU) and medical language processing (MLP) are often used interchangeably, reflecting a similar conceptual overlap to that between Natural Language Understanding (NLU) and natural language processing (NLP). Defining clear boundaries between language understanding and language processing tasks is challenging, as both aim to process and interpret human language across a continuum of complexity from low-level processes such as tokenization and part-of-speech tagging to higher-level tasks like named entity recognition, relation extraction, and temporal event classification (Jurafsky and Martin, 2023). The medical domain, however, adds further layers of complexity due to its abbreviation-heavy vocabulary, inconsistent documentation practices, and the frequent need for contextual and domain-specific knowledge to accurately disambiguate meaning.

This thesis focuses on the evaluation of medical language models, with two primary aims. First, it seeks to understand the current capabilities and limitations of existing models in addressing the complex challenges outlined above. Through comprehensive analysis, we aim to uncover the subtle and non-trivial linguistic nuances that these models continue to struggle with. Second, and more centrally, the thesis pursues two core objectives: (a) to explore the integration of multiple information sources through the development and analysis of multi-sourced medical models as a means to address existing challenges; and (b) to present a reflective study about the proficiency gap between medical language models and human medical experts.

To establish the necessary background for this thesis, the following section outlines key concepts in NLP. We then highlight the distinctive characteristics of medical NLP which shape the assumptions and methodologies adopted in the subsequent chapters.

### 2.1.1 NLP in a Nutshell

The history of human communication over the centuries provides clear evidence of the central role language has played in shaping societies and its continuous evolution. From the pictorial markings of early cave dwellers to the vast collections of inscriptions left by historians, and finally to today's omnipresent source of information — the Internet

[1]This situation is analogous to challenges faced in low-resource language settings, where limited labeled data and domain-specific vocabulary hinders the development of robust language models, however this is not the main interest pursued in this thesis.

— language has remained the cornerstone of knowledge transmission. For much of history, the propagation of information across generations has relied heavily on written texts, ranging from the simplicity of a grandmother's cookbook to the sophistication of Hippocrates' medical treatises[2] (see Figure 2.1). Regardless of form or purpose, these works share a common vision: that written language serves as a powerful medium to preserve, disseminate, and democratize knowledge.

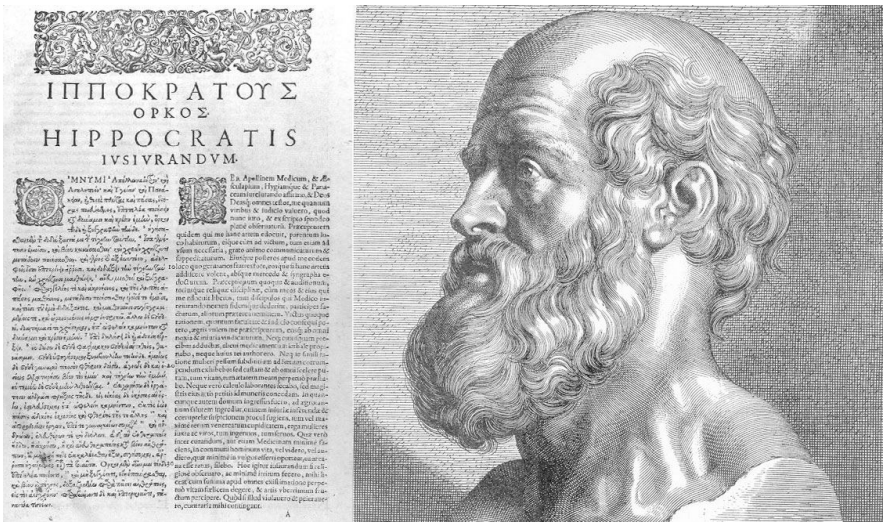



Figure 2.1: Engraving of Hippocrates (right) with an early printed Hippocratic Oath in Greek and Latin (left).

The technological surge brought about by the digital revolution, particularly the advent of personal computers and the internet in the 1980s — granted the general public unprecedented access to vast amounts of information. This era also spurred large-scale efforts to digitize and archive literature across diverse domains. However, as digital content rapidly expanded, traditional search techniques such as basic keyword matching or fuzzy string comparison proved increasingly inadequate (Salton, 1989). The need for more intelligent and efficient methods of navigating large text corpora became apparent, catalyzing early research in information extraction. One of the earliest notable contributions was the "Statistical Machine" a document search engine developed by Emmanuel Goldberg in the early 20th century (Goldberg, 1931). Though primitive by today's standards, this system laid the conceptual foundation for how machines could index, retrieve, and ultimately understand human language, marking a pivotal step in the evolution of automated text processing.

The diversity in the characteristics of text available from different sources in terms of number of languages existing in the world inspired researchers to study the formalism and symbolic nature towards a universal generalization of how language works. This began shaping theoretical models of language structure, providing the foundational frameworks that later led to automatic and more sophisticated approaches for language understanding.

[2]Source: https://www.koulliasgroup.com/blog/480/hippocrates-and-the-aspirin/

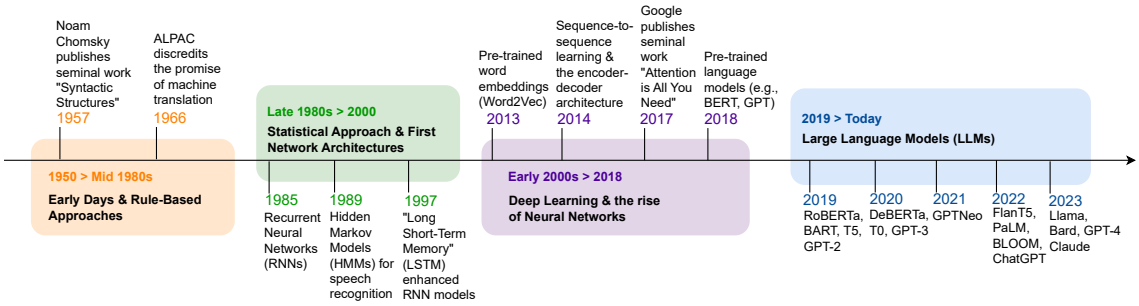



Figure 2.2: Evolution of Natural Language Processing

The evolution of NLP (See fig. 2.2) traces a clear trajectory from early token-level feature engineering to the development of high-level distributional semantics through word embeddings. With the rapid advancement of computational infrastructure, this progression has culminated in the widespread deployment of large language models (LLMs) capable of generating and interpreting human-like language at scale. The core motivation driving this evolution has been the pursuit of more effective and scalable methods for automating information extraction and learning latent representations from raw, unstructured text. These methods increasingly capture complex linguistic patterns and semantic relationships often surpassing the limits of human interpretability.

The simplest way to represent a text excerpt is by a *bag of words*, which can later be used to form co-occurrence matrix. This is formalized as a feature-encoding approach called one-hot encoding where the text can be represented by a vector $v$. The size of the vector $v$ is equal to the assumed vocabulary size $|V|$, here $V$ is the vocabulary. There are different count-based embeddings approaches of encoding features proposed for eg., Tf-idf feature encoding (Salton and Buckley, 1988). Tf-idf feature vector is a normalized vector calculated using term frequency and document frequency. This trick helped to bring down the count effect of stopwords in the text. However, its limitation is in its *sequence invariant* nature i.e. the syntactic order of words is not preserved for eg. “The patient is recovering" and the sentence “recovering is patient the" will have the same representation. Additionally, the fixed length of vocabulary $|V|$ can easily lead to a model failing to generalize to unseen words in spite of being high-dimensional.

Based on the principles of distributional semantics, the introduction of continuous representations in particular word embeddings became a change point in NLP. Word2Vec (Mikolov et al., 2013a,b) marked a significant milestone in the development of pre-trained word embeddings by introducing low-dimensional continuous vector representations that capture semantic and syntactic properties of words as compared to count-based feature embeddings[3]. This approach provided strong evidence for word-level conceptual grounding, as words with similar meanings tend to cluster together in the embedding space. The semantic relationships between words are preserved in these representations, as demonstrated by relational similarity patterns (Pennington

[3]An important advantage is gained over features such as one-hot or tf-idf vector whose length could be as long as the vocabulary $|V| \approx\sim 20K$.

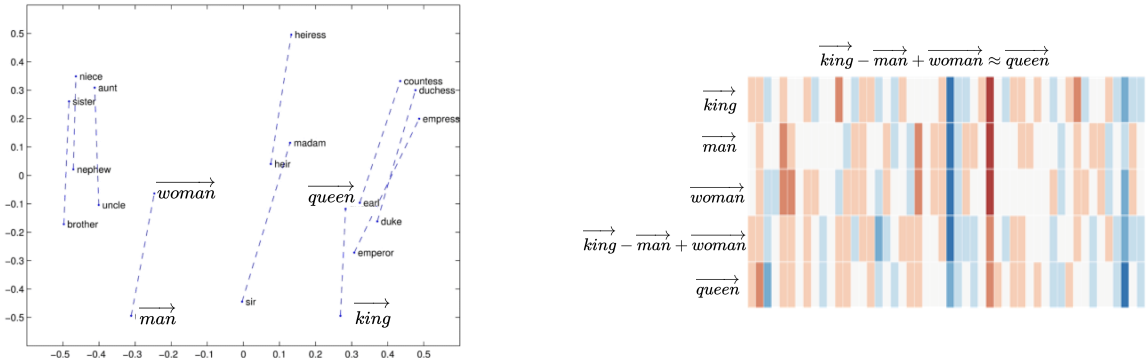



Figure 2.3: GloVe vector space showcasing relational properties.

et al., 2014). A classic example is shown in fig. 2.3, where vector arithmetic such as $\overrightarrow{\text{king}} - \overrightarrow{\text{man}} + \overrightarrow{\text{woman}}$ results in a vector close to $\overrightarrow{\text{queen}}$, illustrating how analogical reasoning can emerge from purely data-driven representations. Here, $\overrightarrow{w}$ denotes the vector representation of a word $w$.

A well-recognized limitation of Word2Vec embeddings lies in their *context-invariant* nature. For instance, the token "queen" would be represented by a single fixed vector irrespective of its usage, whether denoting a royal female monarch or referring to the British rock band. This arises because Word2Vec and similar distributional approaches (e.g., Global Vectors for Word Representation (GloVe); Pennington et al. (2014)) produce a static lookup table of embeddings, learned from global co-occurrence statistics over large corpora. Although such embeddings encode rich semantic and syntactic information, the representational scheme permits only one vector per vocabulary item, thereby disregarding polysemy and contextual variability[4]. In practical natural language understanding tasks, this context insensitivity leads to significant shortcomings, particularly in domains where ambiguity and semantic nuance are critical.

To overcome this drawback, the research focus shifted toward *contextualized word representations*, in which embeddings are dynamically generated as a function of the entire sentence. A pivotal advancement in this direction was the introduction of the Embeddings from Language Models (ELMo) model (Peters et al., 2018), which employs a deep bidirectional language model based on stacked Long Short-Term Memory (LSTM)[5]; Hochreiter and Schmidhuber (1997)) to compute embeddings conditioned on surrounding context. Unlike static embeddings, ELMo representations vary across different occurrences of the same word, allowing the model to capture both polysemy and subtle context-dependent phenomena.

[4]Polysemy refers to a single word having multiple meanings (e.g., "bank" can mean a financial institution or the side of a river), while contextual variability highlights how a word's meaning can change depending on the surrounding words. Static embeddings assign one fixed meaning per word, ignoring these differences.

[5]Long Short-Term Memory (LSTMs) networks are a type of Recurrent Neural Network (RNN) architecture designed to capture long-range dependencies by using gating mechanisms that regulate the flow of information and mitigate the vanishing gradient problem Hochreiter and Schmidhuber, 1997.

An equally crucial development during this period was the refinement of *tokenization*[6] strategies, which directly impacted the quality of learned representations. Earlier approaches typically relied on word-level tokenization, leading to large vocabularies and an inability to effectively handle rare or out-of-vocabulary words. Character-level models offered partial relief by modeling text at finer granularity, as demonstrated by ELMo (Peters et al., 2018). It highlighted the utility of composing meaning from subword-level information, a realization that spurred the adoption of more flexible subword segmentation algorithms. Methods like WordPiece (Wu et al., 2016), Byte-Pair Encoding (Sennrich et al., 2016), and SentencePiece (Kudo and Richardson, 2018) became foundational in scaling models across diverse vocabularies and languages.

Despite these advances, the shift from static to dynamic embeddings was far from instantaneous. Transitional models like Universal Language Model Fine-tuning (ULMFiT) (Howard and Ruder, 2018) demonstrated the value of fine-tuning pre-trained language models on downstream tasks, a technique that became a standard practice. It proposed a three-stage training paradigm consisting of (1) pretraining a general-domain language model, (2) fine-tuning it on a target domain corpus, and (3) adapting it to specific downstream tasks. This approach substantially improved performance on low-resource and domain-specific NLP tasks by enabling models to retain broad linguistic knowledge while adapting to new contexts with minimal labeled data. ULMFiT's success laid the groundwork for large-scale pretraining and task-specific fine-tuning strategies that became foundational to subsequent Transformer-based architectures such as Bidirectional Encoder Representations from Transformers (BERT) and Generative Pretrained Transformer (GPT).

Soon after, a seminal advancement was introduced through the *self-attention mechanism* (Vaswani et al., 2017), which substantially improved the representational capacity of language models. This innovation marked the transition from traditional pretrained language models to the Transformer architecture, enabling models to capture long-range dependencies more effectively and scale efficiently with data and compute. The self-attention formulation allows each token to attend to all other tokens in the sequence, thereby removing the sequential bottleneck inherent in recurrent models such as LSTMs and Gated Recurrent Unit (GRU) (Cho et al., 2014b). As a result, Transformers not only parallelize training but also provide richer hierarchical contextualization, where dependencies are modeled at multiple levels of granu-

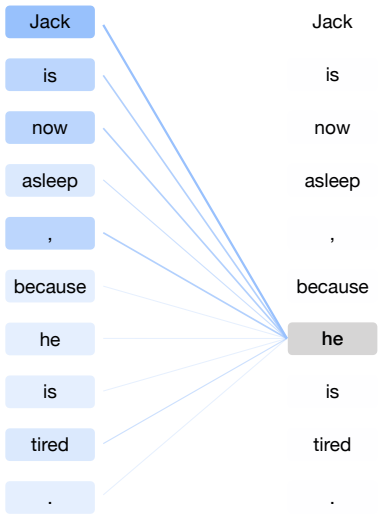



Figure 2.4: An illustration of the self-attention mechanism of Transformer (Han et al., 2021).

[6]Tokenization is the process of breaking down text into smaller units such as words, subwords, or characters that a model can process. The choice of tokenization strategy affects how meaning is captured and learned. We will revisit tokenization in detail in §2.1.2.

larity. Subsequent work demonstrated that stacking deep layers of Transformer blocks further enhanced generalization, establishing the architecture as the foundation of modern NLP.

With growing evidence of the expressive power of such models, the focus of the NLP and computational linguistics community began to shift toward exploring higher-order linguistic phenomena, including inference, reasoning, and text generation. This culminated in the development of large-scale pre-trained Transformers such as BERT (Devlin et al., 2019), which introduced bidirectional masked language modeling, and GPT (Radford et al., 2018), which leveraged auto-regressive training for generative tasks. Unlike GPT's autoregressive, left-to-right formulation, BERT enabled the model to attend to both left and right context simultaneously, resulting in significantly richer contextual representations. Its success across a wide range of tasks with a unified architecture and minimal task-specific adaptation firmly established the pretrain-then-finetune paradigm and demonstrated the power of bidirectional Transformer encoders in modern NLP. These models not only achieved state-of-the-art results on a wide range of benchmarks but also revealed emergent capabilities beyond pattern recognition, motivating a new research agenda around the reasoning and generative potential of large language models.

### 2.1.2 From NLP to Medical NLP

Having introduced the evolution of NLP, we now revisit some of the key concepts with detail to build a more intuitive understanding and bridge them with medical NLP as it will guide the subsequent discussions in this thesis.

**Tokenization.** Modern language models, particularly Transformer-based architectures such as BERT (Devlin et al., 2019) and its variants, employ subword tokenization strategies that decompose words into constituent tokens rather than treating them as atomic units (Sennrich et al., 2016). For instance, the medical term 'adenocarcinoma' is segmented by the vanilla BERT-base model using WordPiece tokenization (Wu et al., 2016) as follows:

$$\text{adenocarcinoma} = \text{aden} + \#\#\text{ocar} + \#\#\text{cin} + \#\#\text{oma} \tag{2.1}$$

where the '##' prefix indicates continuation tokens that are merged with preceding subwords during reconstruction. This subword tokenization paradigm fundamentally differs from traditional word-based embedding approaches such as Word2Vec (Mikolov et al., 2013a) and GloVe (Pennington et al., 2014), which assign a single, fixed-dimensional vector representation to each complete word in the vocabulary. In these conventional models, 'adenocarcinoma' would be represented by a single embedding vector $\mathbf{e}_{\text{adenocarcinoma}} \in \mathbb{R}^d$, where $d$ is the size of the vector. In contrast, modern subword-based models construct the final representation of 'adenocarcinoma' by combining the contextual embeddings of its constituent tokens:

$$\mathbf{h}_{\text{adenocarcinoma}} = f(\mathbf{h}_{\text{aden}}, \mathbf{h}_{\#\#\text{ocar}}, \mathbf{h}_{\#\#\text{cin}}, \mathbf{h}_{\#\#\text{oma}}) \tag{2.2}$$

where $f(\cdot)$ represents the contextual composition function learned by the Transformer architecture, and each $\mathbf{h}_i$ denotes the contextualized representation of the $i$-th subword token.

This tokenization strategy offers several advantages in medical NLP: (1) it enables handling out-of-vocabulary terms through subword composition, (2) it captures morphological regularities in medical terminology (e.g., common prefixes like 'cardio-' or suffixes like '-itis'), and (3) it provides more robust representations for rare medical terms that may not appear frequently in training corpora (Kudo and Richardson, 2018). However, it is important to note that the decomposition of the word adenocarcinoma does not follow classic prefix and suffix used in medicine. And therefore, it also introduces challenges, as the semantic integrity of medical terms may be compromised when decomposed into linguistically unmotivated subword units (Bostrom and Durrett, 2020).

**Embeddings.** Embeddings constitute the foundational paradigm for transforming discrete linguistic tokens into continuous vector representations, serving as the critical interface between symbolic language and computational processing in natural language processing systems. At their core, embeddings provide dense, low-dimensional numerical representations that capture semantic, syntactic, and contextual relationships between words, subwords, or larger linguistic units, enabling machine learning algorithms to operate on textual data through mathematical operations in vector spaces (Mikolov et al., 2013b; Pennington et al., 2014). Formally, an embedding function is defined as

$$\mathbb{E} : \mathcal{V} \rightarrow \mathbb{R}^d \tag{2.3}$$

where $\mathbb{E}$ represents the embedding mapping, $\mathcal{V}$ denotes the vocabulary of discrete tokens, and $d$ specifies the dimensionality of the resulting vector space. This transformation process involves an encoder function that maps raw textual input through learned parameters to produce dense vector representations, where semantically similar tokens are positioned in proximity within the embedding space according to some distance metric, typically cosine similarity or Euclidean distance (Rogers et al., 2021). Embedding techniques has evolved from static approaches like Word2Vec and GloVe, which provided fixed representations regardless of context, to dynamic contextualized embeddings such as ELMo, BERT, and their variants, which generate token representations that vary based on surrounding linguistic context (Peters et al., 2018; Devlin et al., 2019).

In medical NLP applications, specialized embeddings trained on domain-specific corpora, such as Bio-Word2Vec and ClinicalBERT embeddings, demonstrate superior performance by capturing medical terminology, clinical abbreviations, and biomedical semantic relationships that are poorly represented in general-domain embeddings (Zhang et al., 2019c; Alsentzer et al., 2019b). The choice of embedding dimensionality $d$ represents a crucial hyperparameter that balances representational capacity with computational efficiency, where higher dimensions can capture more nuanced semantic distinctions but increase memory requirements and potential overfitting, while lower dimensions provide computational efficiency at the cost of representational fidelity (Yin

and Shen, 2018).

**Sequence Modeling techniques.** The neural network era marked a paradigmatic shift in natural language processing, as the research community recognized the fundamental sequential nature of language and pivoted toward sophisticated sequence modeling architectures capable of capturing temporal dependencies and contextual relationships inherent to textual data. This transition was necessary because of the limitations of traditional bag-of-words approaches and the need to preserve the sequential semantic information where each linguistic element depends critically on preceding context, leading to the development of recurrent neural networks (RNNs), long short-term memory networks (LSTMs), gated recurrent units (GRUs), and ultimately transformer architectures (Hochreiter and Schmidhuber, 1997; Cho et al., 2014a; Vaswani et al., 2017). RNNs, while pioneering in their ability to maintain hidden states across time steps, suffer from vanishing gradient problems and limited capacity to capture long-range dependencies. These issues are addressed by RNN variants such as LSTM and GRU, which use gating mechanisms to selectively retain or forget information over extended sequences (Bengio et al., 1994; Pascanu et al., 2013). The transformer revolution, initiated by the seminal "Attention is All You Need" paper, fundamentally transformed sequence modeling by replacing recurrent architectures with self-attention mechanisms, enabling parallel processing of sequential data while maintaining superior performance on long-range dependency tasks (Vaswani et al., 2017). the self-attention mechanism not only eliminates the limitations of recurrent architectures but also establishes a scalable computational paradigm.

These properties have made it foundational for contemporary large language models, such as GPT and their successors, which leverage massive-scale pre-training on diverse textual corpora followed by task-specific fine-tuning to achieve unprecedented performance across numerous NLP benchmarks (Wang et al., 2018a; Radford et al., 2018; Brown et al., 2020). Contemporary LLMs, including GPT-4, Pathways Language Model (PaLM), and LLaMA, demonstrate promising capabilities in few-shot learning, in-context learning, and reasoning abilities through scaling laws that correlate model performance with parameter count and training data volume (Achiam et al., 2023; Chowdhery et al., 2023; Touvron et al., 2023). Despite their impressive general-domain performance, these sequence modeling techniques encounter significant challenges when applied directly to *medical data* due to the unique characteristics of medical language and lack of biomedical knowledge literature, which we now turn to examine in detail.

## 2.2 Medical Data

Medical data originates from diverse sources and stakeholders within social ecosystems, each contributing unique perspectives and information types. As illustrated in Figure 2.5, the medical data lifecycle encompasses multiple interconnected stages and participants.

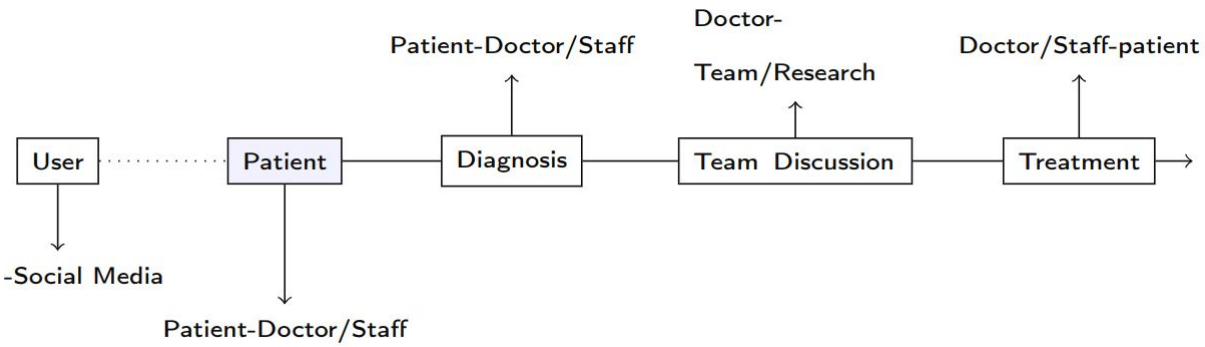



Figure 2.5: Origin of Medical Data.

General public contribute information through social media platforms that may contain health-related discussions and experiences. In the form of patients, they generate data through interactions with healthcare providers during diagnosis and treatment phases. Healthcare professionals, including doctors and staff, create clinical documentation throughout the patient journey, from initial consultations through diagnosis, team discussions, and treatment decisions. These team discussions often involve multidisciplinary research teams who collaborate to determine optimal care strategies. The complexity of medical data is further amplified by its multimodal nature, ranging from structured clinical records and diagnostic reports to unstructured physician notes and patient narratives. This heterogeneous data landscape presents both opportunities and challenges for natural language processing applications, as each data source carries distinct linguistic characteristics, privacy considerations, and clinical relevance that must be carefully addressed in computational approaches to healthcare.

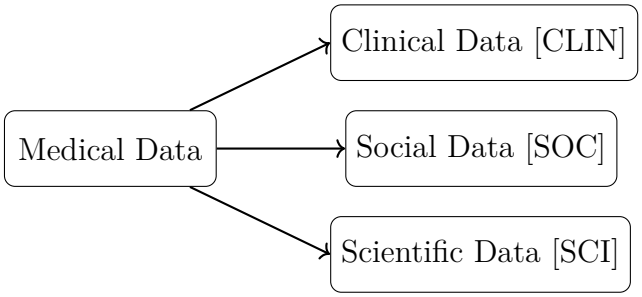



Figure 2.6: Taxonomy of Medical Data

### 2.2.1 Types of Medical Data

We categorize the above described data sources based on their origin into three categories, establishing a comprehensive taxonomy (See Figure 2.6). This will serve as a guide to better explore the behavior of medical language models in a systematic manner and design evaluation strategies:

**Social Media Data [SOC].** Social media platforms have emerged as valuable repositories of health-related discourse, providing insights into patient experiences, public health trends, and disease awareness (Paul et al., 2016; Correia et al., 2020). Social media data presents unique challenges including informal language, abbreviations, misspellings, and privacy concerns, while offering unprecedented access to patient-reported outcomes and real-world experiences (Ginn et al., 2014). This category encompasses:

- **Twitter/X data**: Real-time health discussions, symptom reporting, and medication experiences expressed in short-form text (Sarker et al., 2015)
- **Reddit communities**: In-depth patient narratives from specialized health subreddits (e.g., r/depression, r/diabetes), offering longitudinal patient journeys (Chancellor et al., 2016)
- **Health forums**: Structured discussions on platforms like PatientsLikeMe[7], WebMD[8] forums, and disease-specific communities.
- **Review platforms**: Patient-generated content on healthcare provider review sites and medication feedback platforms (Austin et al., 2020)

**Clinical Data [CLIN].** Clinical data represents the cornerstone of medical NLP applications, comprising structured and unstructured information generated during routine healthcare delivery (Jensen et al., 2012; Wang et al., 2018b). Clinical data is characterized by domain-specific terminology, standardized coding systems (ICD-10, SNOMED-CT), and strict regulatory requirements for privacy and security (Meystre et al., 2008). This encompasses:

- **Electronic Health Records (EHRs)**: Comprehensive patient records including clinical notes, discharge summaries, progress notes, and nursing documentation (Murdoch and Detsky, 2013)
- **Radiology reports**: Detailed interpretations of medical imaging studies, rich in anatomical and pathological descriptions (Pons et al., 2016)
- **Pathology reports**: Microscopic examination findings, tumor staging, and histopathological diagnoses (Spasić et al., 2014)
- **Laboratory data**: Structured test results with associated reference ranges and clinical interpretations (Botsis et al., 2010)
- **Medication records**: Prescription data, dosage information, and administration logs (Uzuner et al., 2010)
- **Clinical trial data**: Structured protocols, adverse event reports, and outcome measurements from controlled studies (Demner-Fushman et al., 2015)

[7] https://www.patientslikeme.com/
[8] https://www.webmd.com/

**Scientific Data [SCI].** Scientific literature constitutes a vast repository of biomedical knowledge, representing decades of research findings and clinical evidence. Scientific data is distinguished by rigorous peer-review processes, standardized formatting, and comprehensive citation networks that enable knowledge graph construction (Westbury et al., 2015). This category includes:

- **Peer-reviewed articles**: Research papers from journals indexed in PubMed, encompassing clinical studies, systematic reviews, and meta-analyses (Lu, 2011)
- **Preprint repositories**: Early-stage research findings from platforms like arXiv, bioRxiv, and medRxiv, providing access to cutting-edge discoveries (Abdill and Blekhman, 2019)
- **Clinical guidelines**: Evidence-based recommendations from professional medical organizations and regulatory bodies (Peleg, 2013)
- **Drug databases**: Comprehensive pharmaceutical information from sources like DrugBank, RxNorm, and FDA drug labels (Law et al., 2014)
- **Biomedical ontologies**: Structured knowledge representations including Gene Ontology, Human Phenotype Ontology, and disease taxonomies (Bodenreider, 2004)
- **Clinical trial registries**: Protocol descriptions and outcome data from ClinicalTrials.gov and similar international repositories (Zarin et al., 2011)

### 2.2.2 Characteristics of Medical Data

Table 2.1 summarizes the key characteristics of three primary categories of medical text sources social (SOC), clinical (CLIN), and scientific (SCI). All three exhibit heterogeneity, reflecting the wide variation in formats, terminologies, and contexts, while also containing unstructured information that poses challenges for automated processing. Social and clinical data are typically multi-modal, combining text with additional signals such as images, metadata, or structured records, whereas scientific texts are largely unimodal, with multi-modality restricted to figures or supplementary content. Temporality is more pronounced in clinical records, where patient histories and longitudinal data are central, and in some types of scientific literature, while social texts tend to lack consistent temporal structure. The three sources also differ in vocabulary richness: social data generally uses a limited and informal lexicon, clinical texts employ a richer specialized vocabulary, and scientific articles are characterized by the most diverse and technical terminology. These contrasts highlight that NLP approaches must be tailored to the specific challenges of each data source, from handling informality and sparsity in social media to managing multimodal and temporal complexity in clinical narratives and coping with highly specialized vocabularies in scientific discourse.

| | Heterogeneity | Multi-modal | Unstructured | Temporal | Vocabulary |
|---|---|---|---|---|---|
| SOC | ✓ | ✓ | ✓ | ×* | ★ |
| CLIN | ✓ | ✓ | ✓ | ✓ | ★★ |
| SCI | ✓ | ×* | ✓ | ×* | ★★★ |

Table 2.1: Characteristics of different medical data sources. * mark implicates that — in trivial setting, the property may not hold, however with a higher-level view the source may exhibits the marked properties as well.

### 2.2.3 Formalization of Medical Data

Medical data encompasses a diverse spectrum of information sources (eg. modalities, sensors-based measurements etc.) that collectively characterize patient health status and clinical scenarios. We formally represent the comprehensive medical data space as $\mathcal{X}$, which captures the heterogeneous nature of information across multiple dimensions. Medical data, in generality, exhibits inherent multi-modality, encompassing various information types that provide complementary perspectives. We can decompose the medical data space $\mathcal{X}$ based on modality $m \in \mathcal{M}$, where $\mathcal{M}$ represents the set of all possible modalities:

$$\mathcal{X} = \{\mathbf{X}^m \mid m \in \mathcal{M}\} \tag{2.4}$$

The principal modalities commonly encountered in medical contexts include:

- **Textual data** ($\mathbf{X}_{text}$): Social Media posts, forum threads, clinical notes, discharge summaries, radiology reports, electronic health records, scientific article and abstracts
- **Images** ($\mathbf{X}_{image}$): Personal photos, Memes, Patient-related images, radiological scans (X-rays, CT, MRI), ultrasound data, illustrations, graph/plots
- **Structured data** ($\mathbf{X}_{tabular}$): Laboratory values, vital signs, demographic information, diagnostic codes and scientific tabular content
- **Physiological signals** ($\mathbf{X}_{signal}$): Electrocardiograms (ECG), electroencephalograms (EEG), and continuous monitoring data

Thus, the complete multi-modal representation can be expressed as:

$$\mathcal{X} = \{\mathbf{X}^{text}, \mathbf{X}^{image}, \mathbf{X}^{tabular}, \mathbf{X}^{signal}, \ldots, \mathbf{X}^{|\mathcal{M}|}\} \tag{2.5}$$

Another fundamental characteristic of medical data to be considered is its temporal nature, reflecting the evolution of information over time (eg. collections of timestamped social media posts or patient condition or scientific articles published over a period of

time). For any given modality $m$, the corresponding data $\mathbf{X}_m$ exhibits time-dependent variations that capture the progression of medical phenomena such as, evolution of topic on social media, patient condition, clinical trajectories or scientific research. We formalize the temporal dimension by introducing a discrete time index $t \in \mathcal{T}$, where $\mathcal{T} = \{t_0, t_1, t_2, \ldots, t_T\}$ represents the ordered sequence of observation timestamps. The temporal unfolding of modality-specific data is represented as:

$$\mathbf{X}^m = \{\mathbf{X}_t^m \mid t \in \mathcal{T}\} = \{\mathbf{X}_{t_0}^m, \mathbf{X}_{t_1}^m, \mathbf{X}_{t_2}^m, \ldots, \mathbf{X}_{t_T}^m\} \tag{2.6}$$

Combining both multi-modal and temporal dimensions, we establish a comprehensive framework for medical data representation. The complete medical data space $\mathcal{X}$ is characterized by the Cartesian product of modalities and temporal indices:

$$\mathcal{X} = \{\mathbf{X}_t^m \mid m \in \mathcal{M}, t \in \mathcal{T}\} \tag{2.7}$$

This unified framework provides the foundation for developing sophisticated medical NLP systems that can effectively process, integrate, and analyze heterogeneous data across multiple dimensions, enabling more comprehensive and contextually-aware clinical decision support.

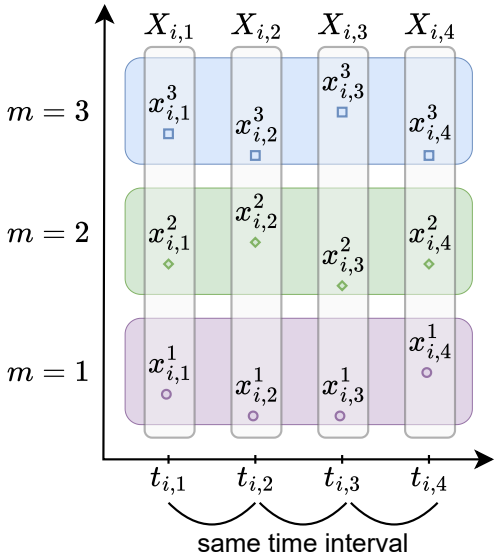



Figure 2.7: Illustration of medical data as a regularly sampled time series.

As an illustrative example of temporal medical data representation, we consider Regularly Sampled Time Series (RSTS). Consider a medical dataset represented by $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$, where $\mathcal{X} = \mathbf{X}_1, \ldots, \mathbf{X}_n$ is the set of instances and $\mathcal{Y} = y_1, \ldots, y_n$ is the corresponding set of labels for $n$ instances. Each instance $\mathbf{X}_i$ represents a time series for the $i^{\text{th}}$ patient, given by $\mathbf{X}i = \{\mathbf{X}_{i,t_1}, \ldots, \mathbf{X}_{i,t_{l_i}}\}$, where $l_i$ denotes the number of time

steps for which measurements were recorded. At each time step $t_j$, $\mathbf{X}i^{t_j}$ contains the values measured across all modalities, i.e., $\mathbf{X}_{i,t_j} = \{x^1_{i,j}, \ldots, x^m_{i,j}\}$, where $x^m_{i,j}$ is the value of modality $m$ for the $i^{\text{th}}$ instance, and $m$ is the total number of modalities[9]. The set of modality indices is denoted by $\mathcal{M} = 1, \ldots, m$. Notably, in RSTS the temporal intervals are uniform, i.e., $t_{j+1} - t_j = \Delta t$ for all consecutive time steps, ensuring regular sampling. While idealized due to their regularity which is not so common in medical data, RSTS provide a clear context to demonstrate how the components of the framework can be integrated. A snapshot of multivariate RSTS data for the $i^{\text{th}}$ instance is shown in fig. 2.7 considering three modalities and four timestamps of measurement.

### 2.2.4 Multi-source Medical Data

Multi-source data encompasses the integration of multiple heterogeneous data streams, representing a fundamental characteristic of ubiquitous raw data. In the medical domain, multi-source data manifests through various combinations of the three primary data modalities: clinical data (electronic health records, diagnostic imaging, laboratory results), scientific data (biomedical literature, clinical guidelines, research findings), and social data (demographic information, socioeconomic factors, patient-reported outcomes) (Bazoge et al., 2023; Kraljevic et al., 2021; Kim et al., 2023). These combinations can take several forms: clinical-scientific integration for evidence-based medical models (He et al., 2022; Gao et al., 2025), clinical-social fusion for population health management (Bazoge et al., 2023), social-scientific synthesis for epidemiological research, or comprehensive tri-modal integration that leverages all three data types simultaneously to provide a holistic understanding of any phenomena (Yang et al., 2024; Kim et al., 2023).

In this context of thesis, we focus primarily on text-based medical data and look into a simpler form of multi-source integration possibilities, where we consider a deep learning model (e.g., a pre-trained language model) in itself as a single source of information, and we investigate the integration of additional data sources such as a clinical knowledge base or additional modality or scientific knowledge graph to operationalize multi-sourced medical data (Gao et al., 2025; He et al., 2022; Zhang et al., 2020).

The label space $\mathcal{Y}$, introduced briefly in RSTS illustration, exhibits considerable diversity depending on the specific medical application: it may be discrete for classification problems, continuous for regression tasks, or sequential for temporal modeling. Throughout this study, we encounter various instantiations of these label spaces across different medical NLP applications. Majority of them may revolve around entity recognition, document classification, language inference, paraphrase generation, and outcome prediction.

Based on the type of problem under study, the notion of input $\mathcal{X}$ can be expanded to different task-specific formulations within medical NLP, as outlined below:

[9]In this context, modality is used interchangeably with feature or sensor. The formulation remains unchanged regardless of the terminology, as any additional source of information (e.g., sensors or derived features) can be incorporated equivalently.

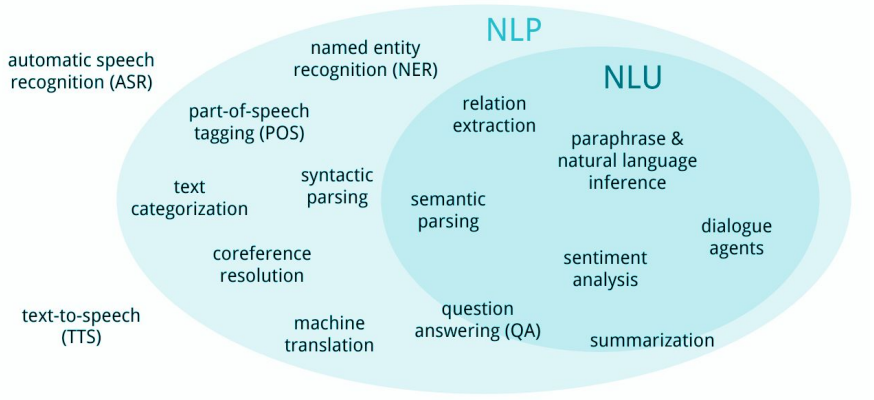



Figure 2.8: Different language tasks that language understanding encompassed in language understanding.

**Entity Recognition.** This is a token-level task in which the language model identifies the semantic class of each token in a given sequence. For example, the model may recognize entities such as medication names, disease mentions, anatomical parts, or procedures. Formally, each token $tkn_i \in \mathcal{X}_\rangle$ is assigned a label $y_i$ corresponding to an entity type, making this a sequence labeling problem. In medical contexts, entity recognition underpins higher-level tasks such as information extraction and concept normalization.

**Language Inference.** This is a sentence-level reasoning task where the model is provided with a pair of sentences a *premise* and a *hypothesis* and must determine their semantic relationship. The relation is typically classified as *entailment*, *contradiction*, or *neutral*, depending on whether the hypothesis logically follows from, contradicts, or is unrelated to the premise. In medical settings, such inference tasks can help in evaluating clinical claims, validating evidence, or assessing consistency between clinical narratives and guidelines.

**Document Classification.** This task involves assigning an entire clinical or biomedical document to one or more predefined categories. Depending on the problem formulation, the number of classes can vary from a few (e.g., identifying report types such as discharge summary or radiology note) to hundreds (e.g., classifying clinical trial abstracts by disease or intervention). Here, the input $\mathcal{X}$ represents the textual content of the scientific document, and the model outputs a categorical label or a probability distribution over possible classes.

**Paraphrase Generation.** This is a generative task in which the model produces a semantically equivalent reformulation of a given input sentence. In medical NLP, paraphrase generation can be related to tasks such as data augmentation, simplification of clinical terminology, and improving robustness of downstream models. For instance, generating alternate phrasings of medical jargon to enables better generaliza-

tion across non-technical crowd or institution-specific linguistic variations or differing documentation styles.

**Outcome Prediction.** This task extends beyond textual understanding to include structured or temporal clinical data. Given a patient's available information such as physiological time-series measurements, laboratory results, or clinical notes collected during an Intensive Care Unit (ICU) stay the model predicts a discrete or continuous clinical outcome. A typical formulation is binary classification, where the goal is to predict patient mortality (0 = survived, 1 = deceased) within a specified observation window (e.g., 48 hours). Outcome prediction tasks exemplify how language models or hybrid architectures can be adapted to handle multimodal medical data for decision support.

## 2.3 When Language Models meets Medical Data

### 2.3.1 Limitations of Language Models

Language models have become foundational tools in modeling and interpreting natural language, offering mechanisms to capture semantic, syntactic, and contextual relationships within a given source of information (Mikolov et al., 2013a; Devlin et al., 2019). However, when applied to the medical domain, models that rely on a single data source be it textual, structured, or temporal encounter inherent limitations in representing the full medical context (Rajkomar et al., 2018b; Shickel et al., 2017). Medical data are inherently multifaceted, encompassing diverse sources such as clinical records (e.g., narratives, laboratory results, imaging) (Dahl et al., 2025; Klug et al., 2024), scientific literature (e.g., recent publications, guidelines, annotated datasets) (Jin et al., 2019a), and social media discourse (e.g., patient-reported experiences, behavioral signals, contextual metadata) (Paul et al., 2016; Bian et al., 2012), each reflecting complementary dimensions of health and contributing to a more comprehensive understanding of medical phenomena. When a model is trained or applied using only one of these sources, it inevitably captures a narrow and incomplete view of the underlying clinical reality (Esteva et al., 2019).

Moreover, medical data are characterized by strong inter-dependencies across different sources, for instance, in the case of multiple modalities — textual reports often reference numerical measurements, while physiological trends may contextualize or contradict diagnostic descriptions (Boag et al., 2018; Harutyunyan et al., 2019). Traditional language-model-based architectures, when constrained to a single modality, lack the capacity to capture cross-source relationships (Huang et al., 2020; Acosta et al., 2022). This leads to gaps in contextual understanding, reduced robustness, and limited transferability across different tasks (Li et al., 2020; Steinberg et al., 2021). Adding to these challenges are, the heterogeneity of medical data, institutional variation, and privacy constraints, all of which exacerbate data sparsity and hinder comprehensive learning from isolated sources (Price and Cohen, 2019; Obermeyer and Emanuel, 2016;

Ghassemi et al., 2020).

Together, these factors underscore the limitations of single-source modeling in the medical domain and motivate the need for frameworks that can effectively integrate multiple, heterogeneous data streams to achieve a more complete and contextually grounded understanding of clinical information (Baltrušaitis et al., 2018; Qiu et al., 2023).

### 2.3.2 Domain Adaptation Challenges

Adapting language models to the medical domain poses substantial challenges arising from the linguistic, structural, and practical characteristics of medical data. These challenges stem from both the inherent differences between general and medical language and the limitations of transfer learning approaches when applied to specialized medical contexts (Devlin et al., 2019; Alsentzer et al., 2019a).

A key obstacle is the linguistic divergence between general-domain and medical text. Medical language is characterized by highly specialized terminology, abbreviations, and lexically precise expressions that convey distinct clinical meanings (Friedman, 1997; Demner-Fushman et al., 2009; Rector and Iannone, 2012). Terms such as *acute*, *chronic*, *benign*, and *malignant* carry context-dependent interpretations, and clinical documentation frequently uses Latin or Greek etymologies, shorthand, and domain-specific acronyms (Beam et al., 2020; Terada et al., 2004; Xu et al., 2010). The syntactic and stylistic structures of clinical notes also differ markedly from natural prose, with fragmented sentences, telegraphic phrasing, and non-standard grammatical constructions optimized for rapid documentation (Meystre et al., 2008; Uzuner et al., 2007; Rosenbloom et al., 2011). Understanding these texts often requires implicit knowledge of specialized workflows and diagnostic reasoning, creating semantic gaps that general-purpose models struggle to bridge (Johnson et al., 2016; Neumann et al., 2019; Peng et al., 2019).

While transfer learning has shown strong performance in general NLP (Devlin et al., 2019; Radford et al., 2018), its effectiveness diminishes in medical contexts (Lee et al., 2020; Huang et al., 2019). Pre-trained models are typically optimized for general corpora and lack the specialized knowledge needed to accurately interpret clinical concepts and relationships (Beltagy et al., 2019; Gu et al., 2021a). Vocabulary mismatches and out-of-vocabulary terms, including abbreviations and rare clinical expressions, further compromise embeddings and semantic understanding (Wu et al., 2015; Zhang et al., 2019c; Lee et al., 2020).

Fine-tuning on medical datasets is also challenging due to data sparsity and accessibility constraints. Annotated clinical corpora are limited in scale and diversity because of privacy regulations, such as HIPAA and GDPR (Price and Cohen, 2019), restricting the ability to train robust, generalizable models. Additionally, institutional and demographics heterogeneity differences in documentation practices, coding standards, and patient populations further limit model transferability across settings (Rajkomar et al., 2019; Obermeyer et al., 2019).

Crucially, these challenges highlight a broader limitation: even when a model is well-adapted to a particular data source, relying on a single source of information be it textual records, laboratory values, or language specific medical discourse provides only a partial view of the medical context. Achieving more comprehensive understanding and robust generalization requires leveraging multiple, complementary sources of information, a perspective that motivates the multi-source integration strategies explored later in this thesis.

### 2.3.3 Role of Knowledge Sources in Medical NLP

To address the limitations of relying on a single source of information, medical natural language processing (NLP) systems increasingly integrate external knowledge sources. These resources provide structured, semi-structured, and curated information that enhances model understanding of clinical language, improves generalization, and supports reasoning over rare or specialized medical concepts (Ji et al., 2021; Petroni et al., 2019; Wang et al., 2021; Zhang et al., 2019d).

**Expert-Curated and Clinical Knowledge.** Expert-curated resources such as clinical guidelines, diagnostic criteria, and treatment protocols provide distilled domain knowledge that complements corpus-based training (Culotta et al., 2006; Shang et al., 2019; Lee et al., 2020; Jin et al., 2019a). These sources are particularly valuable for rare conditions or specialized procedures that are underrepresented in text corpora. Expert annotations and consensus-driven datasets further ensure that models capture clinical reasoning patterns while maintaining validity across diverse healthcare contexts (Uzuner et al., 2007; Kim et al., 2003; Roberts et al., 2009; Pradhan et al., 2013).

**Ontologies and Structured Knowledge.** Among structured resources, ontologies formalize the conceptual structure of a domain, encoding hierarchical relationships, semantic constraints, and logical axioms (Gruber, 1993). In clinical setting, resources such as SNOMED CT, Unified Medical Language System (UMLS), and ICD-10 provide standardized taxonomies of diseases, procedures, and clinical concepts (Bodenreider, 2004; Donnelly et al., 2006; Aronson, 2001). Similarly, for scientific setting, resources such as PubMED, ESMO provide latest literature about clinical trials and practice, and research citation networks (Canese and Weis, 2013; Gennari et al., 2021; Kilicoglu et al., 2012). Integrating ontological knowledge enables models to leverage established clinical semantics rather than relying solely on statistical patterns from textual corpora. Techniques such as knowledge graph embeddings and graph neural networks allow neural models to capture these relationships while maintaining compatibility with transformer-based architectures (Wang et al., 2019; Peters et al., 2019; Liu et al., 2020).

**Medical Modalities.** Medical data by itself offers diverse modalities, such as clinical data which includes narratives, lab results, imaging, and physiological signals that

reflect patient health (Johnson et al., 2016; Zhang et al., 2023a). Scientific data encompass biomedical literature, clinical trial reports, ontologies, and curated knowledge bases, providing structured evidence and domain knowledge (Rebholz-Schuhmann et al., 2012; Lai et al., 2023). Social media data consist of patient-generated content like posts and symptom descriptions, offering real-world insights but posing challenges due to informal language and noise (Chew and Eysenbach, 2010; Paul et al., 2016). Integrating these heterogeneous modalities through multimodal fusion and knowledge graphs improves model robustness and supports comprehensive understanding across medical contexts (Soenksen et al., 2022; Hayat et al., 2022).

**Retrieval-Based and Knowledge-Grounded Approaches.** Retrieval-augmented generation (RAG) frameworks combine parametric knowledge encoded in model weights with non-parametric retrieval from external knowledge bases (Lewis et al., 2020b; Karpukhin et al., 2020; Guu et al., 2020; Borgeaud et al., 2022). Dense retrieval systems identify relevant knowledge for a given query, while generation models synthesize outputs that integrate both retrieved information and learned representations (Karpukhin et al., 2020; Izacard and Grave, 2021). Knowledge grounding mechanisms, including entity linking and fact verification, ensure generated text aligns with authoritative sources, supporting accuracy and reliability in clinical decision-making (Logan et al., 2019; Petroni et al., 2020; Thorne et al., 2018).

Collectively, these knowledge sources can enhance the ability of models to interpret medical language and make informed predictions to aid in mitigating the limitations of single-source reliance. However, most current approaches remain primarily focused on textual or symbolic knowledge, leaving the broader potential for multi-source integration combining complementary patient-level or structured data largely untapped. This observation motivates the exploration of knowledge integration strategies in the subsequent section.

### 2.3.4 Knowledge Integration in Medical Models

The integration of external medical knowledge into language models has emerged as a crucial direction in advancing domain-specific natural language processing. Knowledge integration aims to bridge the gap between purely data-driven learning and the structured reasoning intrinsic to clinical expertise. This is achieved by instilling domain knowledge ranging from terminologies and ontologies to patient-level data and biomedical literature directly into model architectures or training processes.

Early approaches focused on augmenting word or concept embeddings with information derived from medical ontologies such as the Unified Medical Language System (UMLS) and SNOMED CT, where entity-level representations were aligned with concept identifiers to enhance semantic understanding (Bodenreider, 2004; Donnelly et al., 2006). Subsequent models incorporated knowledge during pre-training or fine-tuning through masked entity modeling, relation prediction, or graph-based representation learning, enabling models to capture relationships among diseases, symptoms, and

treatments more effectively. Frameworks such as BioBERT (Lee et al., 2020), PubMedBERT (Gu et al., 2021b) for english language or BioCamemBERT (Touchent et al., 2023), BioMistral (Labrak et al., 2024) and DrBERT (Labrak et al., 2023a) for french among others exemplify this paradigm by integrating domain-specific corpora (e.g., PubMed, MIMIC-III) into the pre-training phase, thereby refining the linguistic and conceptual grounding of the models (Lee et al., 2020; Alsentzer et al., 2019a; Johnson et al., 2016).

More recent work explores knowledge-aware attention mechanisms and cross-modal fusion layers that explicitly condition text representations on structured knowledge sources (Yasunaga et al., 2022; Neil et al., 2018; Jin et al., 2021). For instance, knowledge graphs such as the Semantic MEDLINE Database (SemMedDB) or the Human Phenotype Ontology have been embedded into transformer architectures through relational attention, allowing the model to capture entity relationships and hierarchical structures (Köhler et al., 2017; Kilicoglu et al., 2012; Piya and Beheshti, 2025; Sousa et al., 2023). These approaches enable models to leverage both textual co-occurrences and medically validated relationships, leading to improved interpretability and clinical relevance (Zhang et al., 2021; Gu et al., 2021a; Shang et al., 2019).

Despite these advancements, most existing knowledge integration frameworks remain confined to the textual modality or symbolic concept spaces. They primarily rely on written medical narratives or biomedical literature, overlooking the rich contextual signals embedded in other data forms such as physiological time series, laboratory test results, imaging data, and sensor measurements. As a result, these models capture only a partial view of the patient's clinical context. Achieving a comprehensive representation of medical knowledge therefore necessitates a shift toward multi-modal and multi-source integration strategies that can align textual understanding with complementary temporal and spatial modalities. This broader perspective forms the foundation for developing unified frameworks capable of reasoning across heterogeneous medical data sources.

Multi-source medical modeling contains the potential to bridge these informational silos by combining and often jointly learning from the heterogeneous data streams. Such integration not only enriches the representational capacity of language-model-based systems but also enables more grounded modeling of medical reasoning, where decisions depend on the synthesis of multiple perspective provided by different data sources. Nevertheless, combining multiple data sources introduces its own challenges, including data harmonization, synchronization across time scales, and the handling of missing or partially aligned information. Overcoming these challenges is critical for developing context-aware models capable of robustly generalizing across clinical settings and patient populations.

## 2.4 Conclusion

This chapter has presented the foundational background underpinning this thesis, encompassing the evolution of medical language processing, the characteristics of medical

data, and the development of language models for domain-specific understanding. The discussion began with an overview of natural language processing (NLP) principles and their adaptation to the medical domain, highlighting the unique linguistic, semantic, and contextual challenges that distinguish medical text from general-domain language. It then examined the nature and structure of medical data, outlining its heterogeneity across modalities (including textual, tabular, temporal, and imaging sources) as well as temporal paradigm, challenges associated with their integration and domain adaptation. The chapter further explored the progression of language models and the incorporation of domain knowledge sources through curated and structured resources, underscoring their growing role in medical NLP applications.

While these developments have significantly advanced the capabilities of medical NLP systems, this review also reveals a key limitation: language models, when restricted to one source of information, cannot fully capture the multifaceted nature of clinical data. Medical understanding inherently emerges from the synthesis of information across multiple sources such as clinical notes, physiological signals, patient related images and their experiences each providing complementary evidence toward accurate medical interpretation. Consequently, relying on a single data modality limits contextual awareness and constrains model generalizability across tasks and domains.

On this premise, the next chapter presents a first-hand empirical evidence systematically identifying the issues in single-source medical models, followed by an exploration of methodological approaches for integration of additional source of medical data within a coherent modeling framework. the thesis concludes with a reflection on the gap that persists between medical models and medical human experts in spite of advances we gathered along the way. By doing so, it aims to move beyond single-sourced limitations and potential of multi-sourced modeling toward a more holistic and real-world grounded understanding of the complexity of medical domain.

# Part II

# What Medical Models Can (and Can't) Do

# Chapter 3

# Single-source Medical Models



This chapter presents a systematic evaluation of single-sourced medical models across varying levels of language complexity. It begins with an overview of single-sourced models (§3.1), followed by a series of case studies addressing token-level challenges, including disease extraction in [SOC] data, medication extraction, and complex terminology processing in [CLIN] data (§3.2). It then examines sentence-level issues (§3.3) such as contextual event extraction and the effects of writing style variation in [CLIN] data, before moving to document-level problems (§3.4) involving document understanding in [SCI] data and data scarcity in [CLIN] data. The chapter concludes with an exploration of temporal challenges (§3.5), specifically analyzing missingness patterns in [CLIN] data. Overall, it provides a structured account of the limitations faced by single-sourced models, from granular token-level issues to complex temporal dynamics across diverse medical data types.

## 3.1 What are Single-sourced Medical models?

Conventional learning mechanism in NLP is directly inspired from machine learning practices where we have access to a single dataset $D$[1]. This dataset usually comprises of $(X_i, Y_i)$ where $i$ is the index of the data sample, $X_i$ denotes the input and $Y_i$ is the gold reference. We refer to $X_i$ as the wholesome representation of input data and avoid any non-trivial temporal or modality notion; similarly for the target: $Y_i$. In order to learn a phenomenon, and in the context of thesis a medical phenomenon, we start from scratch with a model $f(\theta)$, where $\theta \in \Theta$ represents randomly chosen initial parameters from $\Theta$ space, such that the model has no prior knowledge about the task or the domain[2]. In such a setting, the only approach which is practiced is to provide the initial model $f(\theta)$ with a subset of the in-hand dataset $D$, referred to as training set, keeping two other unseen subsets namely, validation set and test set which are used during the development and the test phase of the model. Using the aforementioned data splits, the model $f(\theta)$ is trained with the objective as follows:

$$\text{Objective} : min \sum_{x \in X} \mathcal{L}(\hat{y}, y|x) \tag{3.1}$$

In the above equation, $\mathcal{L}$ denotes a loss function or objective function, which penalizes the model based on difference/comparison between $y$ gold reference and the model's prediction $(\hat{y})$. The simplest example of loss function can be difference between the prediction $\hat{y}$ and gold reference $y$, i.e., $|y - \hat{y}|$. This loss function[3] is further subjected to an optimization function such as gradient descent (Sutskever et al., 2013). The optimization function allows the objective function to iteratively overtime attain a global or local minimum which in practice is seen as a plateau in the training loss curve.

There are two important points to note. Firstly, the data splits i.e., training, validation and test set are assumed to come from the same distribution which is also known as Independent and Identically Distributed (IID) assumption (Daume III and Marcu, 2006). And secondly, the conventional objective with which the model $f(\theta)$ is trained is targeted to optimize the models' parameters $(\theta)$ with respect to model's performance given the gold reference $y$ i.e., there is no account of label variability (Plank, 2022; Squires et al., 2023).

Medical language processing, similar to natural language processing, involves problems which range from low-level (eg. surface words) to high-level (eg. semantic space). Therefore, in this chapter, we explore the different challenges that medical language models trained on a single data source face[4] – starting with issues at the token-level.

[1] This can be thought of as an equivalent of knowledge base or knowledge prior.

[2] This can also be thought of as equivalent to a random baseline model. For example, such a model might predict uniformly at random or produce outputs that are statistically indistinguishable from noise, reflecting the absence of learned structure or domain knowledge.

[3] According to the loss function literature, there are several sophisticated options for the optimization function, among which AdamW (Loshchilov and Hutter, 2017) is the most commonly used.

[4] We will refer to them as *single-sourced* medical models.

Research Question 1

What are the challenges faced by Single-Sourced Medical Models?

## 3.2 Token-level problems

Language processing often begins at the token level, where each word or subword unit is analyzed to determine its role or type. In the context of the thesis, this is also a foundational step for understanding medical terminology that would enable the development of more complex semantic and document-level tasks. Accurate token-level representations provide the basis for downstream applications such as entity recognition, relation extraction, and concept understanding. This process is akin to early layers in Convolutional Neural Network (CNN) for facial recognition, where simple patterns like edges are detected before complex features emerge. Similarly, token-level analysis captures basic linguistic units that support higher-level reasoning.

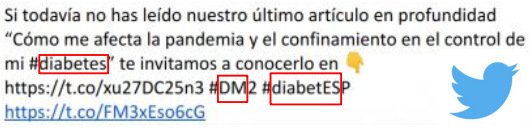



Figure 3.1: Medical data example from social media. Disease mentions (for eg. diabetes, DM) are denoted by □.

In general-domain NLP, token-level classification tasks often focus on identifying named entities such as persons (PER), organizations (ORG), or locations (LOC). In the medical domain, the emphasis shifts to clinically relevant categories such as diseases (DIS), medications (MED), symptoms (SYM), and procedures (PROC). These token-level classifications are typically addressed as Named Entity Recognition (NER)) task (Kundeti et al., 2016), which aims to identify and categorize spans of text that refer to medical concepts.

However, medical NER is considerably more challenging than its general-domain counterpart. Medical terminology is highly variable, frequently abbreviated, and often ambiguous. Entity boundaries may be unclear, and clinical notes or social media posts can contain non-standard spellings, shorthand, or overlapping entities. For example, as shown in Figure 3.1, a user mentions in their tweet in Spanish language a medical condition diabetes using ("diabetESP") that is a non standard spelling which although contains the disease mention yet, highlighting the need for robust token-level models that can handle noisy input.

The quality of token-level understanding directly impacts the reliability of higher-level tasks. Errors at this stage can propagate, undermining performance in tasks such as coreference resolution, temporal reasoning, or document classification.

Therefore, in the following case studies, we explore the interaction of different data sources with *single-sourced* medical language models to operationalize the below research question:

**RQ 1.1**

What are the challenges medical models face on token-level?

### 3.2.1 Case Study I: Disease identification in tweets

This chapter is based on work previously published in our article: *Sinha, A., Holgado, C. G., Clausel, M., & Constant, M. (2022). IAI@ SocialDisNER: Catch me if you can! Capturing complex disease mentions in tweets. In Mining for Health Applications, Workshop & Shared Task (# SMM4H 2022) (p. 85).* Parts of the text, figures, and results are adapted from this publication.

Social media platforms such as Twitter, Reddit, and other public forums with over 4.2 billion users worldwide, have collectively emerged as key sources of information for specific communities. These platforms have become important venues for individuals to share personal opinions and experiences about medications and diseases. From an medical NLP perspective, these platforms offer a rich source of medical data. In order to extract such information, we consider one of the key components in NLP system that is named entity recognition (NER). This component enables us to identify any biomedical entity of interest such as a disease or medication names which can be crucial for downstream applications such as pharmacovigilance, misinformation detection, and the incorporation of patient-reported outcomes into healthcare service design (Murphy et al., 2023; Yu and Vydiswaran, 2022).

Despite state-of-the art models or sophisticated previous proposed methods for general NER benchmark (Yadav and Bethard, 2018), the performance can be affected in the context of entities from Out of Vocabulary (OOV) and suffer from the challenges associated with specialized domain due to its raw data characteristics. More precisely, identifying disease mentions in text from social media can be particularly challenging due to informal spelling, non-standard grammar, frequent use of abbreviations, and the creative linguistic strategies typical of online discourse. For this case study, we focus us on investigating the capability of different language (embedding) models[5] and we frame our research question as follows:

**Research Question 1.1a:** *How well do different embedding models adapt to non standard vocabulary found in [SOC] data?*

## Methodology

**Dataset.** We consider the SocialDisNER dataset (Sánchez et al., 2022) which is a corpus of Spanish Twitter posts annotated for disease mentions. Table 3.1 shows samples illustrating twitter post containing mention of diseases (eg. diabetes and DM2). Posts in this dataset originate from a diverse range of users: (i) patients reporting firsthand health experiences, (ii) friends, relatives, and members of support networks sharing the challenges faced by patients, and (iii) medical professionals disseminating authoritative information about diseases.

[5]We use the term 'language (embedding) model' or 'language model' loosely as an umbrella term to refer to the different models used for word encoding representation which includes static embeddings or contextual embeddings or transformer based embeddings.

| ID | Input | Label | Start | End | Span |
|---|---|---|---|---|---|
| 25 | Si todavía no has leído nuestro último artículo en profundidad "Cómo me afecta la pandemia y el confinamiento en el control de mi #diabetes" te invitamos a conocerlo en https://t.co/xu27DC25n3 #DM2 #diabetESP | ENFERMEDAD (en translation: DISEASE) | 131 | 139 | diabetes |
| | | ENFERMEDAD | 198 | 201 | DM2 |

Table 3.1: Example from SocialDisNER dataset.

Further, the disease mentions in SocialDisNER dataset cover a wide spectrum, including rheumatic diseases such as lupus erythematosus, highly prevalent conditions such as cancer, diabetes, and obesity, as well as mental health disorders, fibromyalgia, and autism spectrum conditions. This diversity highlights the need for token-level medical language understanding capability that can generalize across domains, linguistic variations, and disease categories. Table 3.2 presents the statistics of the SocialDisNER dataset.

| Statistic | Training | Validation |
|---|---|---|
| # Tweets | 5000 | 2500 |
| # Characters | 1,253,431 | 516,768 |
| # Tokens | 211,555 | 84,478 |
| Avg. Characters / Tweet | 250.69 | 206.71 |
| Avg. Tokens / Tweet | 42.31 | 33.79 |
| # Disease Mentions | 15,173 | 4,252 |
| # Unique Disease Mentions | 4,407 | 1,413 |

Table 3.2: Statistics of the SocialDisNER dataset.

**Experimental Setup.** Firstly, minimal preprocessing was performed to preserve as much textual information as possible, as we are interested in investigating the ability of **lm**s to handle this unique characteristic of user-generated data. Tweets were tokenized using whitespace and converted into the CoNLL format (Sang and De Meulder, 2003). For each token, character-based spans were generated, and BIO labels (O= other tag, B-DIS = ENFERMEDAD/Disease) were assigned by aligning tokens with the gold-standard disease mentions provided in the dataset (See Figure 3.2 for an illustration).

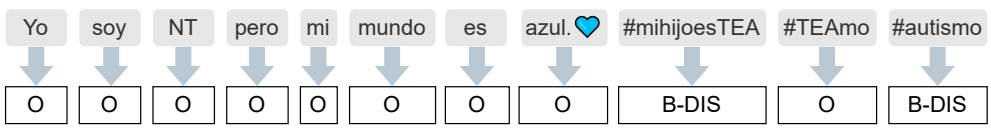



Figure 3.2: CONLL format conversion.

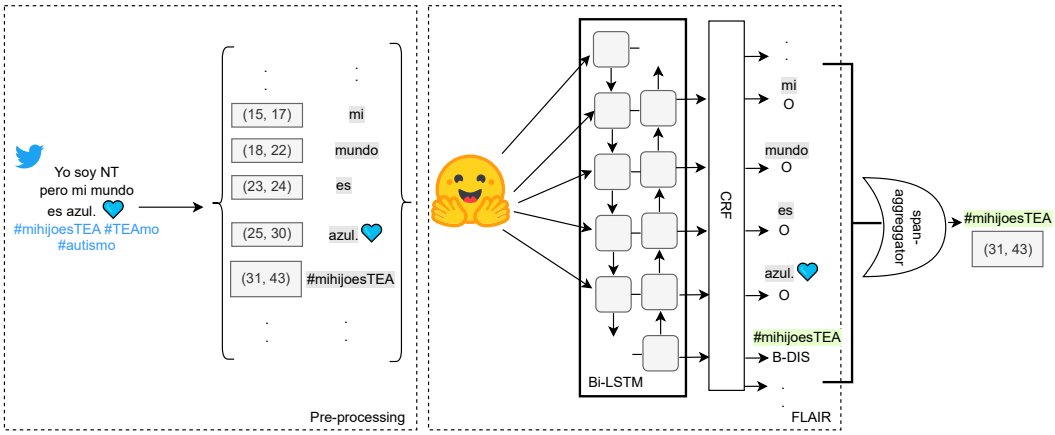



Figure 3.3: Baseline NER Pipeline

To compare the effectiveness of embedding models, we consider the following two dimensions: (i) static vs. contextual representations, and (ii) language- and domain-specific pre-training. This is because we intend to examine the ability of different embedding models to leverage knowledge from related or unrelated domains (Gururangan et al., 2020) and languages (Pfeiffer et al., 2020). For this, we adopt the Fast, Lightweight and Accurate Textual Entailment Recognizer (FLAIR)-NER framework (Akbik et al., 2019), which enables testing a diverse set of embeddings for disease mention extraction. Figure 3.3 presents the illustration of our experimental pipeline.

**Models.** We broadly classify the language embedding models (Language Model (LM)s) employed in our experiments into two groups: *static* and *contextual*. Each group is further organized into subcategories according to their language coverage (`en`, `es`, `multilingual`) and domain specificity (`general` vs. `domain`). For the static embeddings, we use models from the FLAIR Simple (FLAIR-S) package (Akbik et al., 2019), including classical es, en configuration[6], and a combined es+en+clinical configuration[7]. For the contextual embeddings, we draw from the FLAIR-Transformer (Flair-T) package, which integrates models available on HuggingFace. We provide the details regarding the HuggingFace ids for all the models used in Appendix B.2. The positioning of each embedding models with respect to the language-specific and domain-specific aspect is presented in Table 3.3 below:

[6]https://flairnlp.github.io/docs/tutorial-embeddings/classic-word-embeddings
[7]https://flairnlp.github.io/docs/tutorial-embeddings/flair-embeddings

| Type | en | es | multilingual | domain |
|---|---|---|---|---|
| **STATIC** | / | ES | / | ES+CLIN<br>ES+EN+CLIN |
| CONTEXTUAL | BBUCN (Rawal, 2021) | BSCFN (Cañete et al., 2020) | XRL (Conneau et al., 2019)<br>BBMCN (Adelani, 2021)<br>WMN (Tedeschi et al., 2021) | XLRSC (Lange et al., 2021)<br>RBBCE (Carrino et al., 2021)<br>SDF (Chizhikova et al., 2022) |

Table 3.3: Grouping of the different language embeddings models

**Evaluation.** We evaluate the language embedding models' predictions against gold annotations using two criteria: disease identification (Tag-F1), and NER metrics i.e. exact matching (Lenient-F1 and Strict-F1). Tag-F1 denotes the performance of models for identifying the BIO-tags. Lenient-F1 counts any prediction with a non-zero span overlap with the gold as correct, whereas Strict-F1 considers only those predictions correct that exactly match the gold reference span. A higher Lenient-F1 therefore reflects the model's ability to identify the correct disease mention, while a higher Strict-F1 indicates its ability to recover the exact gold-standard boundaries of the mention in the tweet.

To illustrate the difference, let's consider the tweet:

*"Yo soy NT pero mi mundo es azul.♡ #mihijoesTEA #TEAmo #autismo"*

The gold annotations label *"TEA"* and *"autismo"* as Disease. Suppose a model predicts only *"autismo"*. Under Strict-F1, which requires exact boundary matches, this prediction would be partially correct: *"autismo"* matches exactly, but *"TEA"* is missed, resulting in a Strict-F1 of 0.5. In contrast, if the system predicted *"TEAmo"* instead of *"TEA"*, Strict-F1 would mark it as incorrect due to boundary mismatch, whereas Lenient-F1 would give partial score because *"TEA"* overlaps with the gold span. This highlights how Strict-F1 emphasizes precise boundary detection, while Lenient-F1 provides a more relaxed measure that can capture partial matches in noisy text.

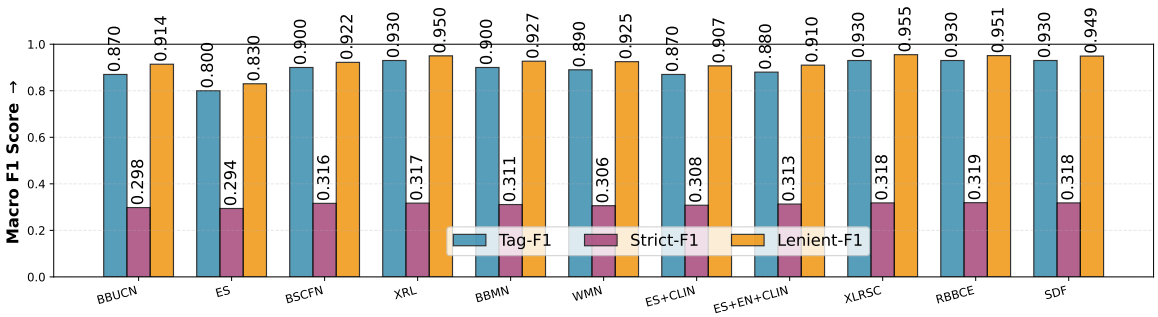



Figure 3.4: Comparison among different language model for disease identification.

## Results & Discussion

Figure 3.4 presents the comparison between different language models for the disease identification task. Table 3.4 shows the details results for identification of BIO Tags i.e. O or B-DIS and corresponding Strict-F1 and Lenient-F1 for all the embedding models for disease extraction NER task in twitter posts. The blue color denotes the contextual embeddings and the red color denotes static embedding models both across the table and the bar plot.

The results demonstrate that contextual language (embedding) models consistently outperform static counterparts on both BIO tag identification and NER evaluation metrics. It is evident that, for Tag-F1 scores — contextual embedding models have an advantage over static models making them more suitable choice for social media (user-generated text) related tasks. Within this group, domain-specific models such as **XLRSC** (Lange et al., 2021), **RBBCE** (Carrino et al., 2021), and Spanish Disease Finder (SDF) (Chizhikova et al., 2022), alongside multilingual models like **BSCFN** (Cañete et al., 2020) and **XRL** (Conneau et al., 2019), achieve highly competitive results with marginal variation.

However, a striking observation is the sharp difference in scores between Tag-F1 and Strict-F1: across all models, Strict-F1 remains below ∼32%. Additionally, for Strict-F1 evaluation, the performance gap between contextual and static embeddings narrows substantially, with differences reduced to around 2%. This contrast suggests that while models are generally capable of detecting the presence of disease mentions, they struggle with identifying the exact token boundaries required for strict span-level evaluation. It is apparent, that predictions often capture relevant spans but include additional character noise, stemming from non-standard spellings, abbreviations, and orthographic variations common in social media content. Further, the consistent low Strict-F1 scores reflect the inherent difficulty of disease mention extraction in noisy social media text.

Addressing this issue likely requires a robust strategies such as post-processing, normalization, or evaluation frameworks that better account for the variability of medical discourse on social media which is not taken care in a trivial single-sourced medical model setting.

## Summary

This case study presents an evaluation of language embedding models for disease identification in social media text. We examine two categories of embedding models: static and contextual. Among these, contextual embeddings consistently outperform static models in identifying the presence of disease mentions at the word level. However, evaluation metrics reveal a notable discrepancy between the Strict-F1 and Lenient-F1 scores. This indicates that, while the models are generally capable of recognizing disease mentions, they often fail to accurately detect the precise token boundaries required

for strict span-level evaluation.

This observation highlights a key *token-level challenge*: although contextual language models can effectively capture the semantic presence of disease mentions, achieving precise span detection in informal, user-generated content remains difficult. This limitation underscores the need for integrating additional knowledge sources or techniques to improve span-level accuracy.

**Limitations** This study does not incorporate Large Language Model (LLM)s in its experimental design, as it was conducted prior to the widespread availability of such models, in the context of the SMM4H-2022 Shared Task[8]. While LLMs may offer improved generalization capabilities for entity detection in user-generated text, the medical domain introduces additional complexities. Exploring the application of LLMs in this context presents a promising direction for future research.

| | BBUCN | ES | BSCFN | XRL | BBMN | WMN | ES+CLIN | ES+EN+CLIN | XLRSC | RBBCE | SDF |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Tag-F1** | 0.87 | 0.80 | 0.90 | 0.93 | 0.90 | 0.89 | 0.87 | 0.88 | 0.93 | 0.93 | 0.93 |
| **Strict-F1** | 0.298 | 0.294 | 0.316 | 0.317 | 0.311 | 0.306 | 0.308 | 0.313 | 0.318 | 0.319 | 0.318 |
| **Lenient-F1** | 0.914 | 0.830 | 0.922 | 0.95 | 0.927 | 0.925 | 0.907 | 0.910 | 0.955 | 0.951 | 0.949 |

Table 3.4: Performance of different language embedding models for disease identification.

[8] https://temu.bsc.es/socialdisner/

### 3.2.2 Case Study II: Medicine extraction in clinical notes

This chapter is based on work previously published in our article: *Sinha, A., Vishwakarma, A., Clausel, M., & Constant, M. (2023). CME² Net: Contextual Medical Event Extraction Network for clinical notes. In CEUR Workshop Proceedings (Vol. 3416, pp. 23-29).* Parts of the text, figures, and results are adapted from this publication.

Clinical notes, documented as part of Electronic Health Record (EHR)s, represent one of the most abundant and richest sources of patient-specific information in clinical setting. Unlike structured fields (e.g., laboratory results or billing codes), clinical notes are predominantly free-text written by clinical practitioners. They capture nuanced details about the patient's symptoms, current medication prescriptions, examination findings, lab/X-ray results, etc (Wang et al., 2015, 2018b; Jensen et al., 2012). Medication information found in these data in particular is critical, as it directly relates to patient safety, treatment adherence, regular assessment and monitoring (Fu et al., 2020). Despite their importance, clinical notes pose substantial challenges for medical language processing as they are inherently *unstructured* and heterogeneous , often containing telegraphic writing styles, domain-specific jargon, abbreviations (e.g., "ASA" for aspirin), and frequent misspellings. (Uzuner et al., 2010; Jagannatha et al., 2019). These factors significantly complicate the extraction of accurate medication mentions.

Pt was a smoker for 50 years occasionally uses **Albuterol** with a spacer suggested he use it pre op . He had a stress test done in May 2094 and it was negative for ischemia and LVEF was 50%. Report in this note under Medical Problems. He was admitted for "chest pain " in May 2094 and he no longer is using the **NTG** or the **Nitropace**. He was told the "pains" he feels is most likely from his CABG ( CABG X3)- Grafts to RCA patent,occluded at circumflex, LIMA patent and the post sternal pain. The stress test was done during this admission as well as a CAT scan to r/o aortic dissection. ( also negative) his biggest problem seems to be his "not very well controlled diabetes". There is a note from the May 2094 admission that he was to receive a trial of **Nexium** to see if his "chest pain" was epigastric" but pt and wife said he never received this medication. He denies any GERD or "chest pain" since may 2094. He recognizes his hypoglycemia sx's and also his hyper glycemia sx's ( blurred vision) when his blood sugar is 480. He will not take his **Glyburide** the night before surgery and the morning of.. He will take 1/2 dose of **Insulin** the morning of 15 untis **NPH**. He denies any medication allergies but stes that "**muscle relaxants**" cause his legs to become flaccid.

Figure 3.5: An excerpt from clinical [CLIN] data from CMED Dataset.

Over the years[9], approaches have evolved from rule-based systems and statistical models (Harkema et al., 2009; Sohn et al., 2010) to neural architectures, including recurrent neural networks and, more recently, transformer-based language models pretrained on biomedical corpora (ValizadehAslani et al., 2023).

In this case study, we focus on medication extraction by framing it as a Named Entity Recognition (NER) task, where the goal is to automatically detect medication mentions within clinical narratives. Figure 3.5 shows an example of clinical note where medication names are denoted in **Bold**, which is often referred to as target medicine.

**Research Question 1.1b:** *How well can embedding based medical language models identify medication names in [CLIN] data ?*

[9]This case study was part of n2c2 challenge 2022 and the related works does not consider large language models.

# Methodology

**Dataset.** To investigate our research question, we utilized the Contextualized Medication Event Dataset (CMED) (Mahajan et al., 2020), which is derived from the 2014 i2b2/UTHealth Natural Language Processing shared task corpus (Kumar et al., 2015; Stubbs et al., 2015). CMED comprises a total of 500 clinical notes, partitioned into 350 for training, 50 for development, and 100 for testing. Table 3.5 presents an example of the medication event annotations provided within the dataset.

| Input | Label | Start | End | Span |
|---|---|---|---|---|
| Pt was a smoker for 50 years occasionally uses Albuterol with a spacer suggested he use it pre op . He had a stress test done in May 2094 and it was negative for ischemia and LVEF was 50in this note under Medical Problems. He was admitted for "chest pain " in May 2094 and he no longer is using the NTG or the Nitropace. He was told the "pains" he feels is most likely from his CABG ( CABG X3)- Grafts to RCA patent,occluded at circumflex, LIMA patent and the post sternal pain. The stress test was done during this admission as well as a CA .... | DRUG | 49 | 57 | Albuterol |
| | DRUG | 310 | 312 | NTG |
| | DRUG | 320 | 328 | Nitropace |

Table 3.5: Example of Medication from CMED dataset.

**Model.** We consider a joint dual-embedding based model architecture, namely Contextual Medical Event Extraction (CME$^2$) network, that utilized an embedding component with access to both static and contextual embeddings representations as a concatenated feature for extracting medications. Figure 3.6 shows the schematic diagram for the CME$^2$ Net model. The complete model contains two heads denoted by Head1 (shown in the figure) and Head2 (removed[10] from the figure). Head1 is complemented with Linear layer that together is utilized for extracting medicine mentions in the clinical note which is the focus for this study.

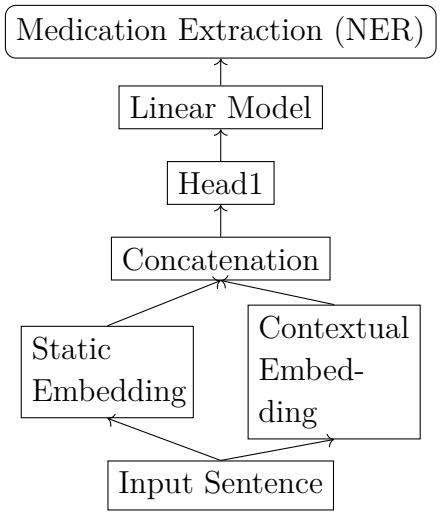



Figure 3.6: Schematic diagram of CME$^2$ Net model.

**Input Preparation.** The embedding component in the CME$^2$ model is designed to combine both static and contextual representations of words. For static embeddings, we utilize pretrained FastText (Joulin et al., 2016) and GloVe embeddings (Pennington et al., 2014), where each word in the input sequence is mapped to its corresponding vector from both

[10]Head2 is removed from Figure 3.6 for readability purpose and it would be introduced in § 3.3.1 (Figure 3.13).

sources, and the two vectors are concatenated. The contextual embedding component consists of three stages. First, a *preprocessor* tokenizes each input word into subword units. Second, the tokenized sequence is passed through a `BERT` layer in a sliding window fashion, producing contextualized embeddings for each subword token. Third, a *postprocessor* merges the subword embeddings corresponding to the same original word by averaging them. For instance, the input word *parkinson* is tokenized into *park*, *##in*, and *##son*; each subword is transformed into a contextual vector, e.g., $[1, 2, 3]$, $[2, 3, 4]$, and $[3, 4, 5]$, and the final word-level embedding is obtained by averaging these vectors, yielding $[2, 3, 4]$. Finally, the static (FastText+GloVe) and contextual (BERT-like) embeddings are concatenated to form a joint representation, which is passed to two prediction heads. The Head1 node is responsible for sequence labeling and consequentially predicts the BIO tagging scheme for Named Entity Recognition.

**Training.** We performed joint training for the CME$^2$ Net Model using an additive weighted loss function $\mathbf{L}$ formulated as follows:

$$\mathbf{L} = \sum_{ij} \mathcal{K}_i \mathcal{L}_j \tag{3.2}$$

In the above equation, $\mathcal{K}_i$ is the softmax weights corresponding to each head in the model. And, for current setup as we are only interested in medication extraction, $\mathcal{K}_2$ was set to 0. Further, each $\mathcal{L}_j$ was a weighted cross entropy loss inspired by Dice Loss (Li et al., 2019b) and is denoted as below:

$$\mathcal{L}_j = \sum_{i=1}^{N} (-y_i \mathtt{ln}(p_i)) * (1 - < p_i >)^{0.1} \tag{3.3}$$

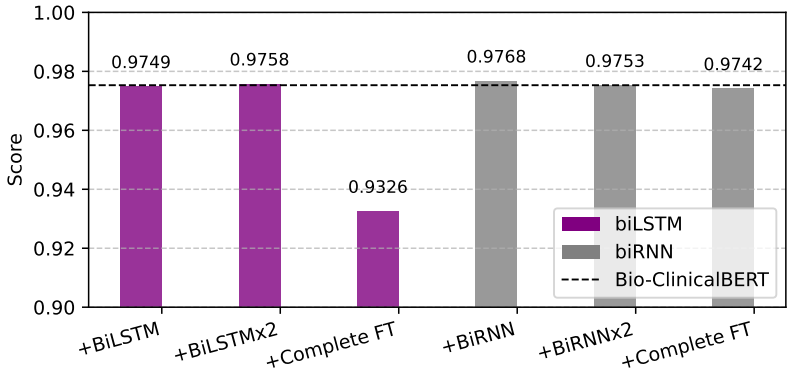
**Hyperparameter Tuning.** Prior to conducting our main experiments, the model architecture initially relied solely on contextualized embeddings (i.e. no static embedding part to start with). We began by selecting a base language model through a comparative evaluation of three pretrained models: BERT-base-cased (Devlin et al., 2019), Bio-ClinicalBERT (Huang et al., 2019), and BioELECTRA (Raj Kanakarajan et al., 2021). Among these, Bio-ClinicalBERT demonstrated the best performance and was subsequently adopted as the embedding backbone for all further experiments.

As part of our hyperparameter tuning, we explored several architectural refinements and regularization strategies, starting from a base model comprising only the Bio-ClinicalBERT encoder for the medication extraction task. Specifically, we experimented with the following configurations:

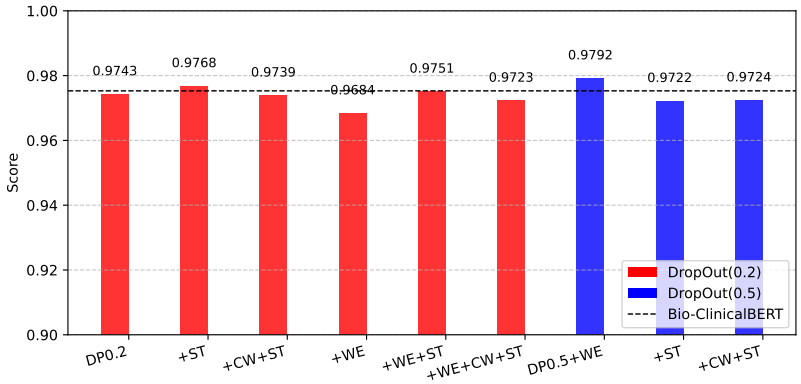
(a) *RNN-LSTM refinement* (Rumeng et al., 2017) : augmenting the model with a Bi-LSTM or Bi-RNN layer stacked above the linear classification head (Figure 3.6) to capture improved latent representations compared to a simple linear decoder (Figure 3.7a);

(b) *Effect of word embeddings (WE), dropout (DP), class weights (CW), and sentence truncation (ST)* : where, in addition to contextualized embeddings, we incorporated static embeddings (e.g., FastText, GloVe), varied the dropout rate {0.2, 0.5},

applied class weighting[11] to address class imbalance, and tested sentence-level truncation strategies (Figure 3.7b).




(a) Effect of RNN-LSTM refinement.




(b) Effect of WordEmbedding (WE), DropOut (DP), ClassWeights (CW), Sentence Truncation (ST).

Figure 3.7: Impact of Hyperparameter tuning on CME$^2$ Model.

For both settings in Figure 3.7, the base model Bio-ClinicalBERT is indicated by the horizontal dashed line in the bar plots at 0.9753 Macro Lenient F1-score. In Figure 3.7a, we observe that additional Bi-LISTMx2 and Bi-RNN to the base model help improve the performance slightly. And, in Figure 3.7b, we observe that addition of word embeddings and dropout rate of 0.5 improves the performance of the model from 0.9753 to 0.9792 Macro Lenient F1-Score.

## Results & Discussion

Figure 3.8 shows summarized performance in comparison to other participants systems and head-to-head comparison with the top 5 models for Lenient-F1 and Strict-F1 met-

[11] Class weights were computed as $1/(\text{frequency}_{\text{class}} + \text{smoothing})$, with a smoothing constant of 100.

rics. In the bar plot (on the right) **Blue** bars denotes Lenient-F1 and **Orange** denotes Strict-F1 evaluation. Our proposed model, CME$^2$ Net, obtained 0.983 for Lenient-F1 (whereas best system obtained 0.985) and 0.9588 for Strict-F1 score (whereas best system obtained 0.9716). CME$^2$ Net model obtained 2nd position out of 32 teams for clinical medication extraction subtask for the n2c2 challenge (Mahajan et al., 2020).

| | Min | Max | CME$^2$ Net |
|---|---|---|---|
| **Strict F1** | 0.0913 | 0.9716 | 0.9588 |
| **Lenient F1** | 0.0943 | 0.9846 | 0.9831 |

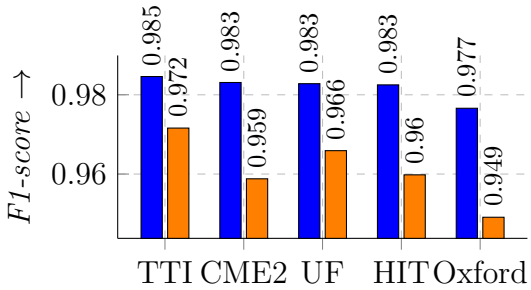



Figure 3.8: Summarized results and head-to-head performance of top 5 models on CMED test set for medication extraction task.

During error analysis, we observed that the CME$^2$ model correctly identified 962 out of 1010 medication names. Closer inspection revealed that most errors were tied to token-level challenges rather than the complete failure to recognize a medication. For example, compound mentions such as "Insulin NPH" or "contrast dye" were frequently split into separate tokens ("Insulin" and "NPH"), leading to partial recognition. Conversely, in cases of agglomeration (e.g., "lipitor20"), the model successfully recovered the span, suggesting robustness to certain character-level variations. Brand names such as "CARDURA" and "Lamictal" were more problematic, as they were consistently missed, pointing to limited vocabulary coverage. Similarly, formulations containing delimiters (e.g., "Lisinopril/HCTZ", "Ca 600/vit D") were often fragmented into multiple entities rather than identified as a single medication span. Inconsistencies were also observed when the same medication appeared multiple times in a document, with the model tagging it differently across mentions.

These findings resonate with observations from the official challenge overview (Mahajan et al., 2023). In general, medications with very short names (e.g., "Ca", "K", "$O_2$", "EPO", "MOM") were missed more frequently than those with longer names (Uzuner et al., 2010). However, exceptions such as "ASA", "NPH", "ARB", "BB", and "NTG" were captured more reliably, likely due to their prevalence in the training corpus, which was enriched for cardiac and diabetic medications, given its derivation from the i2b2/UTHealth risk factor corpus.

Overall, these errors highlight a central *token-level* problem: while the model can often detect fragments of a medication mention, it struggles to consistently capture complete spans when faced with short forms, compounding, delimiters, brand names, or repeated mentions. This underscores the importance of designing evaluation and modeling strategies that explicitly account for tokenization and span-boundary sensitivity in clinical medication extraction.

## Summary

In this case study, we expanded our focus from disease identification in social media ([SOC]) data to medication extraction from clinical ([CLIN]) data. We proposed a dual-embedding model architecture, named *Contextual Medical Event Extraction* (CME$^2$ Net), which integrates clinical static embeddings and ClinicalBERT. The model was trained using a weighted cross-entropy loss inspired by Dice loss to better handle class imbalance. CME$^2$ Net was submitted to the 2022 n2c2 shared task and achieved an overall second place ranking out of 32 participating teams. The our proposed model demonstrated strong performance in extracting medication names; however, it struggled with abbreviations and the variability in clinical writing styles, which remains a challenge in this domain.

**Limitations** As with our previous study, large language models (LLMs) were not included in this work due to the timing of the study, and their effectiveness for this task remains an open question for future research. Additionally, a methodological difference exists between this and our earlier disease extraction study: the previous work employed the FLAIR framework a user-friendly, plug-and-play library while this study utilized the more modular CME$^2$ Net architecture.

Unlike FLAIR framework (Akbik et al., 2019), CME$^2$ Net allows for greater flexibility and fine-tuning of individual components, which enabled more targeted performance optimization. This contrast underscores the potential benefits of modular design in medical NLP systems, offering enhanced control and adaptability for domain-specific challenges.

### 3.2.3 Case Study III: Complex terminologies

This chapter is based on work previously published in our article: *Buhnila, I., Sinha, A., & Constant, M. (2024, August). Retrieve, Generate, Evaluate: A Case Study for Medical Paraphrases Generation with Small Language Models. In Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024), pages 189–203, Bangkok, Thailand. Association for Computational Linguistics.* Parts of the text, figures, and results are adapted from this publication.

Domain-specific vocabulary is a fundamental challenge when working with medical language processing. Clinical narratives and biomedical discourse, or related interactions with clinicians frequently include scientific disease names or chemical compound terms that are unfamiliar to non-experts and this unfamiliarity also is applicable to general domain language models. The vocabulary mismatch is widely recognized as a root source of error in medical NLP systems, due to the high term variation and rare or specialized tokens encountered in medical notes (Uzuner et al., 2010; Leaman et al., 2015). This issue motivated the development of domain-specific models and vocabulary adaptation techniques (Balde et al., 2025; Kwon et al., 2022).

In real clinician–patient communication, clinicians often simplify or explain specialized terminology in order to ensure patient comprehension and engagement. This linguistic process aligns with the NLP task of *text simplification*, which is particularly important in medical domains. Medical text simplification systems aim to either replace or explain them in ways that make clinical content more accessible (Kandula et al., 2010; Leaman et al., 2015; Llanos et al., 2016a).

At the token level, unfamiliar domain-specific words are problematic because model then must rely on subword decomposition and internal knowledge. General domain LMs may break down a rare or novel medical term into subwords and assemble meaning from them, but this process is error-prone when terms are highly specialized or deviate from common morphological patterns (See examples in Table 3.6).

| LMs | "akathisia" | "hypopnea" |
|---|---|---|
| BERT (Devlin et al., 2019) | aka + this + ia | hypo + pne + a |
| BioBERT (Lee et al., 2020) | akathisia | hypopnea |
| LLaMA-3-3B (AI@Meta, 2024) | aka + this + ia | hypo + pnea |
| BioMistral-7B (Labrak et al., 2024) | akathisia | hypopnea |

Table 3.6: Comparison of subword decomposition for medical terms across general-domain and domain-specific models.

In this case study, we examine language models to generate definitions of medical terms, with the additional requirement that definitions remain as simple as possible in order to guarantee that the model does not just understand rare tokens but to produce human-friendly descriptions in a constrained lexicon.

**Research Question 1.1c:** *How well can language models generate simple definitions for [SCI] medical terms?*

# Methodology

**Dataset.** We consider a collection of French medical terms and corresponding sub-sentential paraphrases named RefoMED dataset (Buhnila and Todirascu, 2023). The dataset was built by automatically extracting sentences from the following source corpora, namely ClassYN (Todirascu et al., 2012) and CLEAR Cochrane (Grabar and Cardon, 2018). The paraphrases were identified with the help of linguistic paraphrase markers such as *c'est-à-dire* ("so called"), *également appelé* ("also called"), *est une maladie* ("is a disease"), and punctuation signs, such as colons and brackets. Table 3.7 shows examples from the RefoMED dataset with corresponding translation in english (**en**). The dataset is made of 6297 pairs of unique medical terms and their corresponding sub-sentential paraphrases.

| | Term | Paraphase |
|---|---|---|
| **fr** | hypopnée | respiration partiellement bloquée |
| **en** | hypopnea | partially blocked breathing |
| **fr** | akathisie | agitation intérieure et incapacité à rester assis |
| **en** | akathisia | inner restlessness and inability to sit still |

Table 3.7: Samples of medical term paraphrase from RefoMED dataset.

**Task Description.** We address the task of medical text simplification, which can also be viewed as a form of paraphrase generation. The goal is to take medical jargon as input and use language models to generate simplified explanations or definitions.

**Models.** We considered two text generation models capable of producing text in French: BARTHEZ[12] (Eddine et al., 2020), an encoder–decoder model trained for natural language generation tasks, and BioMistral-7B (Labrak et al., 2024), a multilingual medical large language model[13].

**Experimental Setup.** To generate definitions from these models, we employed two approaches: (i) vanilla prompting and (ii) additional instruction fine-tuning[14]. For both models, we used the following zero-shot vanilla prompting with the prompt shown in Figure 3.21. For instruction fine-tuning, we used used the Quantized Low-Rank Adaptation (Q-LoRA) (Dettmers et al., 2024) method, a computationally efficient finetuning

[12]We used BARThezOrangeSum-abstract model configuration, but for readability reasons, we will refer to BARThezOrangeSum-abstract as BARTHEZ.

[13]We specifically used the `BioMistral-7B-SLERP` configuration, as it achieved the best benchmark performance on French datasets according to Labrak et al. (2024).

[14]For performing instruction fine-tuning the models, we split the dataset by unique term entry while staying in the range of the classic 60-20-20 percentage split for train-val-test sets. The resulting split comprised of 3981 term-paraphrase pairs for training, 1063 for validation and 1253 pairs for testing.

method that reduces the number of parameters for BioMistral from 7B to 1.38B parameters. The prompt for instruction finetuning the models is provided in Figure 3.10 in which the base prompt is the same as vanilla prompt with additional formatting for indicating instruction ([INST] . . . [/INST]) along with gold reference paraphrase.

For both the models, with all the configurations we perform 2 sets of experiments, one where we allow only 25 new tokens to be generated by the model (referred to as *Tokens=25*) and the other where the models is allowed to generate 50 new tokens ( referred to as *Tokens=50*). These settings enable us to test the fluency of the language models based on shortness of the generated paraphrase.

**Evaluation.** We evaluate the generated definitions using two types of metrics: automatic text generation metrics and manual quality-based metrics. The automatic metrics include: Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004), which measures $n$-gram overlap with an emphasis on recall; Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002), which computes the precision of $n$-gram matches between the output and the reference; BERTScore (Zhang et al., 2019a), which compares token embeddings between the output and reference text; and Bilingual Evaluation Understudy with Representations from Transformers (BLEURT) (Sellam et al., 2020), which estimates semantic similarity while accounting for lexical differences. For manual evaluation, we consider a set of the 1200 examples that was analyzed by 3 French proficient linguist annotators following criteria : (a) *readibility*; (b) *completeness*; and (c) *correctness*. Table 3.8 provides the description of each of the above mentioned criteria for automatic and manual evaluation.

| Criterion | Description |
|---|---|
| ROUGE (↑) | Measures n-gram overlap with the reference text, emphasizing recall. Scores range from 0 to 1 (or 0%–100%). |
| BLEU (↑) | Measures n-gram precision between the candidate and reference texts. Scores range from 0 to 1 (or 0%–100%). |
| BERTScore (↑) | Assesses semantic similarity using contextual embeddings (token-level cosine similarity). Typical range is around 0.0 to 1.0. |
| BLEURT (↑) | Uses a learned model to estimate semantic similarity, accounting for lexical variation. Scores usually range from around -1.0 to 1.0 or higher. |
| *Readability* (↓) | Scored from 1 to 3: (1) fluent, grammatically correct, and easy to understand for laypeople; (2) includes invented words, English words, or grammatical mistakes, or scientific terms used correctly but in complex contexts; (3) combines the issues of score 2 with dense scientific terminology, making it difficult for laypeople to understand. |
| *Completeness* (↑) | Indicates whether the generated text provides a full and concise answer (score 1 if complete, 0 otherwise). Two types: *relaxed*—text contains one incomplete sentence or a second incomplete clause; *strict*—text contains at least one syntactically independent sentence. |
| *Correctness* (↑) | Measures whether the text contains correct medical knowledge and is written in French (score 1 if both conditions met, 0 otherwise). Two types: *relaxed*—the general meaning of the medical term is understandable; *strict*—the exact meaning is both understandable and complete. |

Table 3.8: Evaluation Criteria for Medical Paraphrase Detection.

# Results & Discussion

Table 3.9 presents the evaluation of BARThez and BIOMISTRAL on the paraphrase generation task using standard automatic metrics (BERTScore, BLEURT, BLEU-1, ROUGE-1) and human evaluation metrics (readability, completeness, correctness) as described in Table 3.8. In the left section ***Tokens=25*** presents results for all the different setting for the two models with and without Fine-Tuning (FT) and the right section i.e. ***Tokens=50*** presents the one-to-one change in result that was observed when models were allowed longer generation, gray color implies no change observed; green color font implies the model's generated paraphrase improved and red implied the model generations degraded based on the respective metric.

For *Tokens=25*, BIOMISTRAL consistently outperforms BARThez across all automatic metrics. With fine-tuning, BIOMISTRAL achieves a BERTScore of 0.72 compared to 0.62 for BARThez, and its ROUGE-1 score reaches 0.22 versus 0.11. BLEURT values are close for both models (∼0.15), but BIOMISTRAL remains slightly higher. Fine-tuning improves both models, although the gains are more pronounced for BIOMISTRAL. When increasing to *Tokens=50*, BARThez remains stable, showing almost no variation (±0.00 to +0.01 across metrics). In contrast, BIOMISTRAL exhibits small degradations in most metrics (between −0.02 and −0.04), indicating that

| Metric | BARThez w/o FT | BARThez w/ FT | BIOMISTRAL w/o FT | BIOMISTRAL w/ FT |
|---|---|---|---|---|
| ***Tokens = 25*** | | | | |
| BERTScore | $0.63_{0.03}$ | $0.62_{0.02}$ | $0.70_{0.06}$ | $0.72_{0.07}$ |
| BLEURT | $0.10_{0.10}$ | $0.06_{0.08}$ | $0.15_{0.15}$ | $0.15_{0.17}$ |
| BLEU-1 | $0.04_{0.06}$ | $0.06_{0.07}$ | $0.11_{0.12}$ | $0.14_{0.13}$ |
| ROUGE-1 | $0.07_{0.08}$ | $0.11_{0.08}$ | $0.20_{0.16}$ | $0.22_{0.17}$ |
| Read. | 1.22 | 1.36 | 1.08 | 1.34 |
| Compl. (St.) | 100 | 0 | 10 | 16 |
| Compl. (Len.) | 100 | 0 | 20 | 20 |
| Correc. (St.) | 0 | 0 | 68 | 90 |
| Correc. (Len.) | 0 | 0 | 96 | 94 |

| Metric | BARThez w/o FT | BARThez w/ FT | BIOMISTRAL w/o FT | BIOMISTRAL w/ FT |
|---|---|---|---|---|
| ***Tokens = 50*** | | | | |
| BERTScore | ±0.00 | +0.01 | -0.02 | -0.03 |
| BLEURT | ±0.00 | +0.04 | +0.01 | +0.01 |
| BLEU-1 | ±0.00 | ±0.00 | -0.03 | -0.04 |
| ROUGE-1 | ±0.00 | +0.01 | -0.02 | -0.04 |
| Read. | -0.02 | +0.06 | +0.02 | +0.16 |
| Compl. (St.) | ±0.00 | ±0.00 | +8 | +8 |
| Compl. (Len.) | ±0.00 | ±0.00 | +76 | +22 |
| Correc. (St.) | ±0.00 | ±0.00 | +26 | ±0.00 |
| Correc. (Len.) | ±0.00 | ±0.00 | ±0.00 | ±0.00 |

Table 3.9: Evaluation of BARThez and BIOMISTRAL across automatic and manual metrics for different token limits.

BIOMISTRAL is more sensitive to token length while BARThez appears more robust. Overall, BIOMISTRAL achieves the highest scores under optimal settings (25 tokens with fine-tuning).

Human evaluation further highlights the differences between the two models. *Readability* is comparable across models (∼1.2–1.3), with slightly higher values for BIOMISTRAL after fine-tuning. For *Completeness*, BARThez shows worse performance—often scoring close to zero after fine-tuning—while BIOMISTRAL achieves much higher values, particularly with longer generations (up to +76 improvement in length completeness at 50 tokens). *Correctness* also favors BIOMISTRAL, with large gains over BARThez, which consistently scores near zero.

These results suggest that BIOMISTRAL is the stronger model for paraphrase generation, especially when fine-tuned with shorter contexts (25 tokens). BARThez, while more stable across different token lengths, consistently lags behind in completeness and correctness. In practical terms, BIOMISTRAL produces more faithful and complete paraphrases, whereas BARThez outputs are simpler but less reliable which also can be attributed to the role of knowledge prior in these language models. BARThez although is more equipped in terms of language compared to BioMistral however, it still lacks access to any biomedical knowledge which is an advantage for BioMistral.

The behavior of BARThez highlights two characteristics of general domain models when applied to medical domain, without access to specialized knowledge and task specificity they will fail to generalize to medical domain.

## Summary

In this case study, we evaluated language models for the problem of medical text simplification with a French paraphrase dataset *RefoMED*. We utilized two pretrained language models adapted to the biomedical domain as followsw: BARThez and BioMistral.

Overall, BARThez, which is originally designed for summarization tasks, struggled

across nearly all configurations, while BioMistral consistently outperformed it. Manual evaluation further revealed that BioMistral produced outputs with improved clinical correctness on fine tuning. The under-performance of BARThez highlights the characteristic of general domain language models and emphasis on the need for integrating additional external knowledge sources to enhance relevance and informativeness in medical text simplification.

**Limitations** It is important to note that the choice of evaluation criteria plays a critical role in the medical domain. Discrepancies often arise between automatic text generation metrics and manual or clinically grounded evaluations, potentially leading to differing interpretations of model performance. In light of recent literature, large language models (LLMs) can also be employed in an *LLM-as-a-judge* paradigm (Gu et al., 2024) to partially address limitations in manual evaluation. However, this approach introduces its own challenges, such as potential bias due to the model evaluating its own outputs.

**fr** Expliquez-moi le terme médical en mots simples, par une paraphrase ou une courte définition :

**en** Explain the medical term to me in simple words, through a paraphrase or a short definition :

Figure 3.9: Our prompt template for inference.

[INST]
Expliquez-moi le terme médical en mots simples, par une paraphrase ou une courte définition : {`term`},
/[/INST]
{`paraphrase`}

Figure 3.10: Our instruction finetuning prompt template.

## 3.3 Sentence-level problems

Sentence-level modeling is another foundational component of medical NLP pipelines. Tasks such as named entity recognition (NER), terminology definition, and paraphrase generation, discussed in §3.2, operate primarily at the token level, but these tasks often rely on sentence-level modeling to resolve broader contextual challenges. Typical downstream applications include context-based classification, negation detection (Chapman et al., 2001; Uzuner et al., 2007), and relation extraction (Li et al., 2016; Shivade et al., 2014), which are commonly framed at the sentence level and serve as critical building blocks for higher-level tasks such as patient phenotyping (Alzoubi et al., 2019), clinical decision support, and information retrieval (Roberts et al., 2016). With the emergence of transformer-based models like BERT (Devlin et al., 2019) and its domain-specific variants (e.g., BioBERT (Lee et al., 2020), ClinicalBERT (Alsentzer et al., 2019b)), performance on many sentence-level benchmarks has improved significantly.

However, medical data in particular clinical narratives present unique challenges for sentence-level modeling. They are often short, fragmented, and irregular, shaped by time constraints, individual clinician writing styles, and institutional conventions (Uzuner et al., 2008). Sentences may contain critical medical information but also exhibit non-standard linguistic phenomena such as heavy use of abbreviations, omitted subjects, and idiosyncratic shorthand, all of which deviate from conventional grammar. As a result, even basic preprocessing steps such as sentence segmentation and tokenization can introduce ambiguity, propagating errors into downstream tasks.

Medical sentences have in general the characteristic of being densely packed with information. A single sentence may encode multiple entities, overlapping relations, and temporally or causally connected events (Sun et al., 2013b). Capturing these nuanced structures requires fine-grained syntactic and semantic understanding, which is not always consistently expressed in clinical text. For instance, determining whether a medication is active, discontinued, or modified (e.g., "Metformin was discontinued last year") requires sentence-level contextual interpretation beyond entity recognition. Such factors underscore why sentence-level processing in the medical domain remains a persistent and complex challenge.

We present a series of case studies examining various medical text data settings to evaluate the capabilities of language models particularly medical language models trained on single data sources and to identify the challenges they encounter in sentence-level tasks. Formally, we pose the following overarching research question:

**RQ 1.2**

What are the challenges medical models face on sentence-level?

### 3.3.1 Case Study IV: Contextual Event Extraction

This chapter is based on work previously published in our article: *Sinha, A., Vishwakarma, A., Clausel, M., & Constant, M. (2023). CME² Net: Contextual Medical Event Extraction Network for clinical notes. In CEUR Workshop Proceedings (Vol. 3416, pp. 23-29).* Parts of the text, figures, and results are adapted from this publication.

Understanding medication changes is crucial for reconstructing a patient's medical history. Figure 3.11 provides an illustration for the different type of context information associated to any medication mention that can be of interest to clinicians. However, as most clinical notes are written in unstructured, narrative formats and often long free form, it is challenging to automatically identify and interpret such events. This task typically requires expert human annotators, a process that can be both time-consuming and costly. To address these challenges, various approaches have been proposed, including machine learning (Sohn et al., 2010; Gkotsis et al., 2016) and deep learning techniques (Rumeng et al., 2017; Bhatia et al., 2018), aimed at extracting medical entities and classifying contextual events. Rule-based systems (Harkema et al., 2009) and traditional machine learning models such as Decision Trees, Naïve Bayes, and SVMs often struggle with out-of-vocabulary terms, imbalanced datasets, indirect expressions of state changes, and subtle variations in event definitions. Meanwhile, deep learning approaches, including hybrid RNN-residual network architectures (Rumeng et al., 2017) and multitask learning frameworks (Bhatia et al., 2018), are more robust but still face challenges. These include ambiguity arising from diverse writing styles in electronic health records (EHRs), such as misspellings, abbreviations, and inconsistent tense usage.

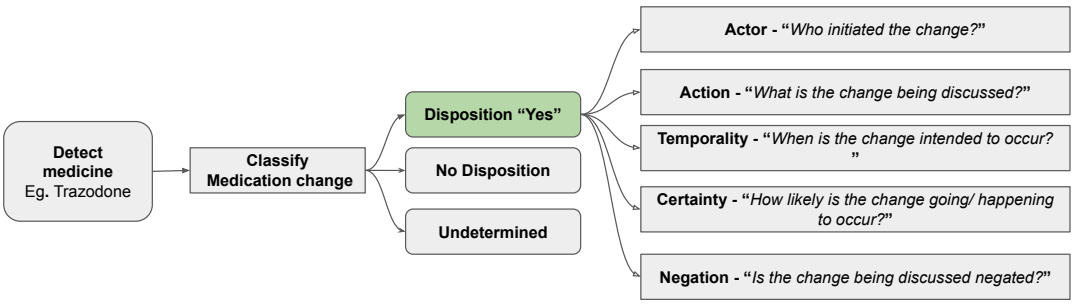



Figure 3.11: Illustration of context associated to medication change.

Given the complexity and variability inherent in clinical narratives, an important open question remains: to what extent can modern language models, particularly those trained on medical data, effectively capture and interpret the nuanced context surrounding medication changes? In this study, we are interested in evaluating language models across diverse scenarios such as this, to better understand their strengths and limitations. Hence, we ask:

**Research Question 1.2a:** *How well can medical LMs capture rich context in [CLIN] medical data?*

# Methodology

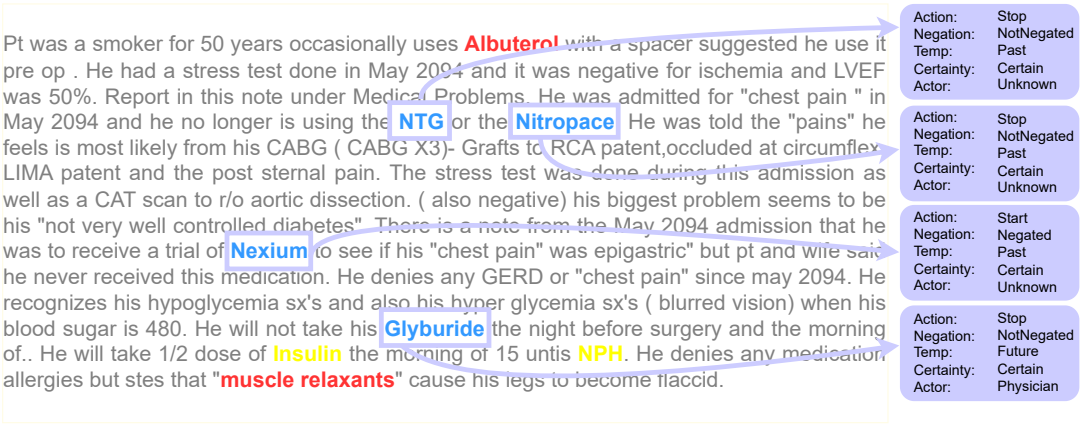



Figure 3.12: Excerpt from clinical data.

**Dataset.** For this case study, we use the CMED dataset (Mahajan et al., 2022) introduced in section §3.2.2. It contains a collection of 500 clinical notes that are fine-grained annotated for medication change events and associated context to the medications. For each clinical note, the dataset contains spans of all the mentioned medications, along with event labels about whether each medication is associated to a *medication_change* referred to as *Disposition* (otherwise *NoDisposition* or *Undetermined*). Figure 3.11 shows illustrations of annotation guidelines and fig. 3.12 shows examples of the medication indicated with colors for events (eg. Disposition, NoDisposition, and Undetermined) along with contextual annotations for each Disposition medication in all context categories. For medication instances with *Disposition* label, we are provided with contextual labels of the medication change event based on the following contextual dimensions: `Action`, `Negation`, `Actor`, `Temporality` and `Certainty`.

**Task Description.** We consider two tasks: (a) event classification and (b) context classification. Given a clinical note and a target medication as input, the language model is required to perform both tasks. First, in the event classification step, the model determines whether a change in the target medication is being discussed. If a change is identified, the model then performs context classification to further categorize the nature of the change. For a detailed illustration, refer to fig. 3.11.

**Model.** We extend the use of the CME$^2$ model which was earlier introduced in §3.2.2 for the following tasks: medical event classification and contextual classification. The CME$^2$ model architecture (See Figure 3.13) consists of embedding component with two encoder modules i.e. static embeddings (FastText, GloVe and Part of Speech (POS)-Tagging-scheme) and a contextual embedding module, a feature concatenation layer,

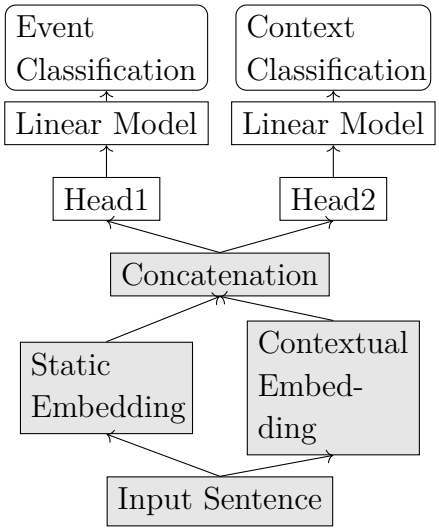



Figure 3.13: CME$^2$ Net Model complete Architecture.

and two separate linear head layers named Head1 and Head2[15].

The complete model undergoes joint end-to-end training with adaptive sample weighting loss which can be denoted by the below equation (borrowed from §3.2.2):

$$\begin{aligned} \mathbf{L}(\mathcal{L}_1, \mathcal{L}_2) &= \mathcal{K}_1\mathcal{L}_1 + \mathcal{K}_2\mathcal{L}_2 \\ \mathcal{L}_j &= \sum_{i=1}^{N}(-\texttt{ln}(p_{i,y_i})) * (1- < p_{i,y_i} >)^{\gamma} \end{aligned} \tag{3.4}$$

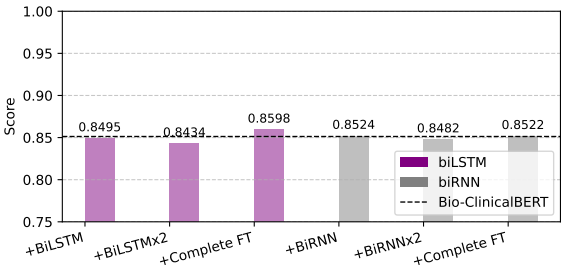
Here, $p_{i,y_i}$ refers to the model's predicted probability for the true class $y_i$. $\gamma \geq 0$ is the focusing parameter that controls how much to down-weight easy examples. The value of $\gamma$ was put to 0.1 for all the experiments. The $K_i$ coefficient in the loss function weights assigned to each Head to perform a weighted joint training of the model.

**Metrics** For both the event classification and context classification task, Lenient Macro F1-score was used for evaluation.

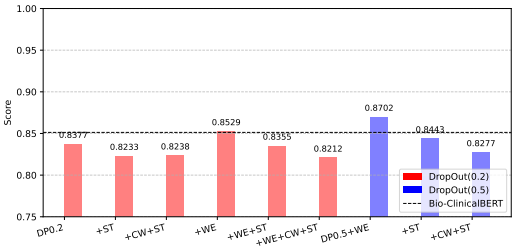
**Hyperparameter Tuning.** The training is done first with assigning $\mathcal{K}_2$=0 in eq. (3.4), when we perform inference for event classification using the model on the validation set under the two criteria of hyperparameter tuning introduced in §3.2.2 namely, a) *RNN-LSTM refinement* (See Figure 3.14a) where an bi-LSTM or bi-RNN is added on the top of Head1 node for refinement; (b) *Effect of word embedding, dropout, class weighting and sentence-truncation* (See Figure 3.14b) in addition to contextualized embeddings, we incorporated static embeddings (e.g., FastText, GloVe), varied the dropout rate

[15]In §3.2.2, we only utilized Head1 for clinical entity extraction (eg. Nitropace was labeled as DRUG) and for this case we have additional information for the drug that it is associated to DISPOSITION as its event label.

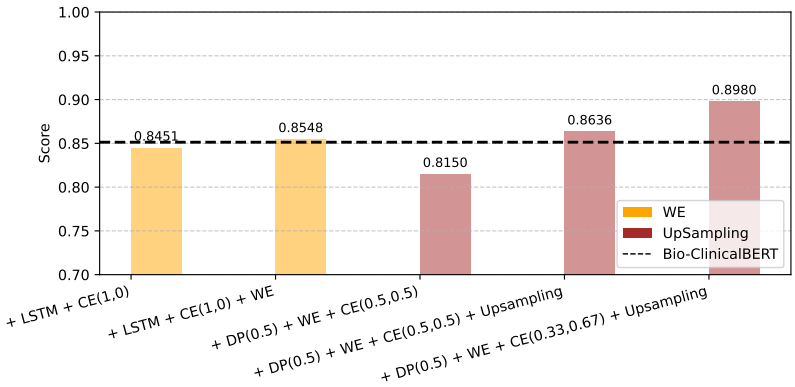
{0.2, 0.5}, applied class weighting[16] to address class imbalance, and tested sentence-level truncation strategies; and later experiment with additional setting with the second classification head Head2 in the model to investigate the (c) *Effect of using bidirectional LSTM layer on word embeddings and Upsampling* (See Figure 3.14c) where we add a bi-LSTM layer on the top of static word embeddings, we also introduce Upsampling during the training to sample more often samples from minority classes (See Figure 3.16) and finally we conducted joint training with both the classification heads using different configurations of CE($K_1$, $K_2$).




(a) Effect of RNN-LSTM refinement.




(b) Effect of WordEmbedding (WE), DropOut (DP), ClassWeights (CW), Sentence Truncation (ST).




(c) Effect of Bi-LSTM layer, Upsampling and Complete Training.

Figure 3.14: Impact of hyperparameter tuning on CME$^2$ Net model components.

# Results & Discussion

Figure 3.14 shows that our base model with Bio-ClinicalBERT with the horizontal dashed line indicating a Macro (lenient) F1-score 0.8513 for contextual event classification. First, when a bi-LSTM layer is added to the base model, the macro F1 for task 2 decreased to 0.8434. Similarly, when bi-RNN layer was added to the base model, the macro F1-score for task 2 decreased to 0.8482. Next, we tried different combinations of dropout {0.2, 0.5}, adding static Word Embedding (WE) layer, assigning Class

[16]Class weights were computed as $1/(\text{frequency}_{\text{class}} + \text{smoothing})$, with a smoothing constant of 100.

Weights (CW) and Sentence Truncation (ST) (see Figure 3.14b). We observed that sentence truncation and using class weights did not improve the results. We then added static word embeddings, for which we used GloVe and FastText embeddings trained on Open Access Case Reports (Flamholz et al., 2022). We observed that adding word embeddings and dropout rate of 0.5 increased the macro F1-score for event classification 0.8702.

| Task | Metric | Max | Min | CME$^2$ Net | Overall Rank |
|---|---|---|---|---|---|
| Event | Lenient micro F1 | 0.9379 | 0.4243 | 0.9272 | 4th/32 |
| | Lenient macro F1 | 0.8673 | 0.2663 | 0.8417 | |
| Context | Combined Lenient F1 | 0.7297 | 0.0209 | 0.6912 | 2nd/32 |

Figure 3.15: Test results for EVENT and CONTEXT classification task.

Adding a biLSTM layer over static word embeddings yielded an $F_1$ score of 0.8451 on Task 2, which did not result in any meaningful improvement. Among all tested configurations built on top of Bio-ClinicalBERT, the best performance was obtained using a dropout rate of 0.5 combined with an additional static word embedding encoder layer. Training with a cross-entropy loss and balanced weighting coefficients $(\mathcal{K}_1, \mathcal{K}_2) = (0.5, 0.5)$ produced $F_1$ scores of 0.8150 and 0.5144 on the event and context classification tasks, respectively. We next trained the model with upsampling (see Figure 3.14), which enhanced performance across both tasks. Finally, retraining with upsampling and adjusted loss weights $(\mathcal{K}_1, \mathcal{K}_2) = (0.33, 0.67)$ further improved performance, achieving $F_1$ scores of 0.8980 and 0.5857 on the two tasks, respectively.

With the gold labels for task 1, our best submission obtained **0.9272** F1-score on event classification for strict matching. Our post-evaluation system obtained **0.6912** F1-score on context classification.

Error analysis revealed that the model correctly classified 167 of 201 event instances. Among the remaining errors, nearly two-thirds resulted from ambiguity between the *Undetermined* and *Disposition/ UnDisposition* classes. This suggests that local sentence context surrounding medication mentions strongly influences classification accuracy, and that subtle linguistic ambiguity often hinders reliable predictions. A closer examination identified 70 misclassifications across the contextual categories. Most involved `Action` (43 cases), followed by `Certainty/Temporality` (17 cases), `Actor` (13 cases), and `Negation` (4 cases). Errors in `Action` frequently arose when medications appeared in lists or long narrative sentences describing patient history, where complex syntax obscured whether an instance should be interpreted as *Start* or *UniqueDose*. Similar ambiguity affected `Certainty`, particularly in hypothetical or conditional statements (e.g., "if symptoms persist, start medication"). For `Actor`, the model often misattributed agency when both *Patient* and *Physician* were mentioned in the preceding context, hindering correct attribution of the action.

Interestingly, the model occasionally succeeded in handling colloquial or abbreviated sentence constructions (e.g., "inc" or "taper off"), even when these were not

aligned with gold annotations. This highlights both the difficulty of *sentence-level* interpretation in clinical narratives and the mismatch between natural clinical phrasing and rigid annotation guidelines (which could also be annotator disagreement) .

## Summary

In this case study, we investigated the tasks of event and context classification related to medication changes in clinical notes. We utilized the CME$^2$ model, previously introduced in Section §3.2.2. This work formed part of the 2022 n2c2 challenge, where the CME$^2$ model demonstrated strong performance ranking fourth in the event classification task and second in the context classification task out of 32 participating teams.

Building upon the earlier medication extraction task described in token-level case study[17], our findings reveal a noticeable performance gap between entity extraction and the context classification task obtained here with a gap of ∼29% between Lenient-F1 Score. This gap reflects an implicit increase in task complexity, shifting from token-level predictions to sentence-level reasoning. Furthermore, the disparity in performance between event classification and context classification suggests additional differences in task difficulty.

Our error analysis indicates variability in model's performance across different context classes, suggesting that the model struggles with certain classes more than others. These challenges appear to be compounded by the variability in clinical writing styles, which may further contribute to the overall complexity of the task. Overall, these findings indicate the language model struggles with rich context found in clinical data.

**Limitations** This study does not incorporate large language models in its experimental design, as it builds directly upon the earlier medication extraction work discussed in Section §section 3.2.2, which was conducted as part of the 2022 n2c2 challenge.

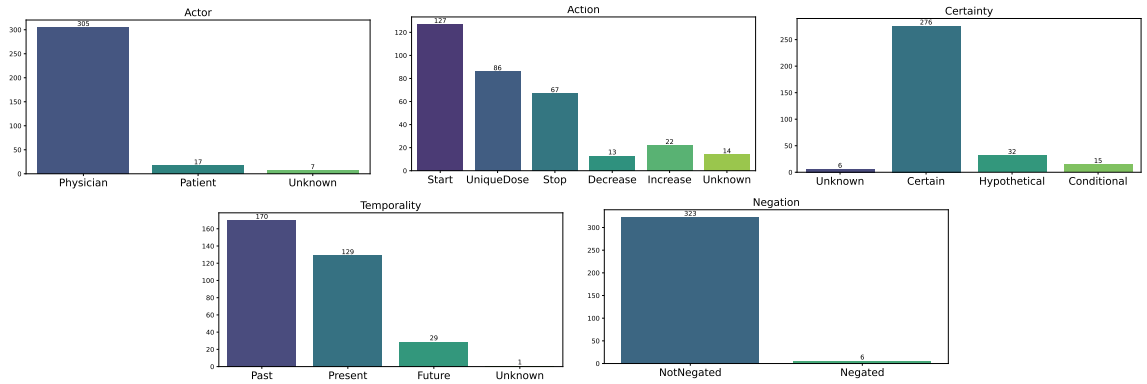



Figure 3.16: Distribution of context labels across CMED trainset.

[17]CME$^2$ Net model achieved overall Strict-F1 Score of 0.9588 on medication extraction task.

### 3.3.2 Case Study V: Impact of writing style variation

The style in which clinical narratives are written plays a critical role in shaping how a patient's history is documented and later interpreted. Beyond providing answers to prognostic inquiries, these narratives are central to disentangling disease progression from the effects of interventions (Pham et al., 2017). Typically, clinical notes follow a chronological order and are structured into recognizable sections such as prior medical history, physical examination, potential diagnoses, and treatment plans (Bansler et al., 2016). However, differences in individual writing styles can significantly affect the clarity and reliability of the record. Variation in narrative style may lead to omissions, ambiguities, or even misrepresentations, which in turn can complicate clinical reasoning and decision recommendation for medical models.

English for Medical Purposes (EMP) has long been a priority in medical education, and several researchers have emphasized the importance of acquiring collocations for English as a Foreign Language (EFL) learners from diverse linguistic backgrounds (Džuganová, 2019). Nevertheless, medical writing is frequently characterized by long and complex sentences, misuse or imprecise use of basic terms, overreliance on passive voice, weak nouns replacing stronger verbs, and excessive use of medical jargon. These stylistic issues can introduce redundancy, ambiguity, and hinder comprehension. For example, Millar and Budgell (2019) found that texts with heavy passive voice or dense medical vocabulary were less comprehensible to health science students compared to versions simplified using more direct language. Such problems underlines why standard and precise writing practices are essential in clinical documentation. Poorly structured text not only affects human readability but also diminishes the reliability of downstream computational analysis.

An illustrative example can be seen in the following sentence (Goodman and Edwards, 2014): *The infusion rate was then increased and blood was taken 4, 8, 12 and 20 minutes after the new target concentration had been achieved.*

The writer intended to describe how quickly the concentration reached a new steady state after the infusion rate was increased. In practice, blood samples were collected 4, 8, 12, and 20 minutes after the adjustment intended to establish the new target concentration that is, after altering the infusion rate. However, the original wording is ambiguous: it suggests that the investigators already knew when the new concentration had been "achieved," whereas in fact the drug levels were not available until laboratory analysis was completed days later.

A recommended alternative (Goodman and Edwards, 2014) is: *The infusion rate was then increased and blood was taken 4, 8, 12 and 20 minutes later, or (although it is implicit and not strictly necessary) . . . after the new target concentration had been set.*

This revision clarifies the sequence of actions and avoids implying knowledge that the investigators did not yet possess. More broadly, this illustrates a common challenge in medical writing: ambiguity often arises when procedural actions and subsequent outcomes are conflated. Distinguishing between what was *done* in real time and what

was *learned* later is essential for precise and unambiguous documentation.

The consequences of writing style variation extend beyond human comprehension. Because unstructured free-text narratives account for nearly 80% of all clinical data, medical models trained on this data are inevitably exposed to stylistic inconsistencies. In practice, the individual writing style of clinicians may influence how models interpret and prioritize information. To examine this, we perform a study investigating if linguistic cues in clinical notes affect model performance and look into the following research question:

**Research Question 1.2b:** *Are medical LMs affected by different writing styles in [CLIN] medical data?*

# Methodology

**Dataset.** To study the above research question, we utilize again the CMED dataset (Mahajan et al., 2022) that we used in some of the previous studies (cf. §3.2.2, §3.3.1). The CMED dataset was constructed using the 2014 i2b2/ UTHealth Natural Language Processing shared task corpus. The clinical notes for this corpus came from the Partners HealthCare Electronic Medical Record (EMR). EMR at Partners HealthCare comprises a platform shared by two large academic tertiary hospitals – Massachusetts General Hospital (MGH) and Brigham and Women's Hospital (BWH), USA. The clinical notes are associated to the patients who were admitted for Type 2 Diabetes (T2D). It was annotated by a team of three annotators led by a physician for medication events extraction and context annotations across different notes. We use the test data and segment the sentences (see example in Table 3.10) that contain medication names to have access to the immediate sentence for each target medication.

| Raw paragraph from Clinical Note |
|---|
| Pt was a smoker for 50 years, occasionally uses Albuterol with a spacer and was suggested to use it pre-op. He had a stress test done in May 2094 and it was negative for ischemia and LVEF was 50%. Report in this note under Medical Problems. He was admitted for "chest pain" in May 2094 and he no longer is using the NTG or the Nitropace. He was told the "pains" he feels is most likely from his CABG (CABG X3) — grafts to RCA patent, occluded at circumflex, LIMA patent and the post-sternal pain. The stress test was done during this admission as well as a CA ... |
| **Segmented Sentence** |
| Pt was a smoker for 50 years, occasionally uses **Albuterol** with a spacer (suggested pre-op). |
| He was admitted for "chest pain" in May 2094 and he no longer is using the **NTG** or the **Nitropace**. |

Table 3.10: Example of Medical Sentence Segmentation.

**Model.** We consider four Pretrained Language Model (PLM) based models, two general domain and two domain specific models : BERT (Devlin et al., 2018), ModernBERT (Warner et al., 2024), ClinicalBERT (Huang et al., 2019) and PubmedBERT (Gu et al., 2021b).

**Experimental Setup.** For this task, we retain all segmented sentences that contain a medication change annotated with a corresponding disposition label. We first employ the pre-trained language models described earlier to generate context-class predictions on the validation and test sets. Next, combine the predictions based on majority voting on sample-level to obtain the ensemble predictions. Sample-level *correctness* is then defined by comparing ensemble predictions with the gold-standard labels. We make use of Building Educational Applications (BEA) linguistic feature extractor tool Lee and Lee (2023) in order to extract linguistic features from the segmented medical sentence of the clinical notes. The tool facilitates to extract for every sentence 220 linguistic feature. These linguistic features cover various linguistic characteristics on sentence-level such as surface, syntax, lexico-semantics, and discourse. We provide the list of all the linguistic features in Tables B.2 and B.3. We utilized SciSpacy[18] additionally instead of general domain modules for parsing the segmented sentences.

To address our research question, we apply the Ordinary Least Squares (OLS) method[19] (Kuchibhotla et al., 2019) to assess the linear correlations between sample-level correctness and the extracted linguistic features, analyzed separately across the five context categories.

## Results & Discussion

We show in Table 3.11 the percentage of linguistic features that were significantly correlated with prediction mistakes in dev and test set for all the five context categories. We provide the list of all the significant correlated linguistic features from Table 3.13 - 3.20. We also provide the list of top five linguistic features that were identified to impact model's correctness most strongly along with broader linguistic categorization for each features (see Table 3.12) based on Lee and Lee (2023)'s documentation which provides the linguistic property of each of the selected features namely, linguistic family and domain.

[18] en_core_sci_sm

[19] Ordinary Least Squares (OLS) is a fundamental regression technique that estimates the relationship between input features and a target variable by fitting a linear model that minimizes the average squared difference between the predicted and true values. In practice, OLS finds the line (or hyperplane) that best fits the data in a least-squares sense, providing interpretable coefficients under standard statistical assumptions.

| | Actor | Action | Certainty | Temporality | Negation |
|---|---|---|---|---|---|
| DEV | 3.64 | 42.27 | 27.73 | 12.27 | 67.73 |
| TEST | 11.82 | 37.73 | 21.36 | 7.73 | 9.55 |

Table 3.11: Percentage (%) of Linguistic Features that are significantly ($\rho$<0.05) correlated with prediction mistakes.

| Action | Actor | Certainty |
|---|---|---|
| avg_brysbaert_age_of_acquisition_of_words (avgworddiff, lex-sem) | #unique_verbs (partofspeech, syntax) | corrected_verbs_variation (lexicalvariation, lex-sem) |
| corrected_adverbs_variation (lexicalvariation, lex-sem) | #named_entities_cardinal (entity, discourse) | root_verbs_variation (lexicalvariation, lex-sem) |
| root_adverbs_variation (avgworddiff, lex-sem) | avg_adverbs_per_word (partofspeech, syntax) | avg_particles_per_word (partofspeech, syntax) |
| #unique_nouns (partofspeech, syntax) | #named_entities_loc (entity, discourse) | avg_adjectives_per_word (partofspeech, syntax) |
| simple_numerals_variation (lexicalvariation, lex-sem) | avg_named_entities_norp_per_sent (entity, discourse) | #unique_nouns (avgworddiff, lex-sem) |

| Negation | Temporality |
|---|---|
| simple_adpositions_variation (avgworddiff, lex-sem) | #unique_punctuations (partofspeech, syntax) |
| #named_entities_product (entity, discourse) | simple_type_token_ratio_no_lemma (typetokenratio, lex-sem) |
| smog_index (readformula, surface) | simple_type_token_ratio (typetokenratio, lex-sem) |
| avg_named_entities_loc_per_word (entity, discourse) | simple_punctuation_variation (lexicalvariation, lex-sem) |
| avg_named_entities_time_per_sent (entity, discourse) | #unique_nouns (partofspeech, syntax) |

Table 3.12: Top 5 significant linguistic features for each context category in the order of decreasing effect size. We indicate the linguistic family in blue and linguistic domain in purple for each feature.

Among the five context categories, *Actor*, *Certainty*, and *Negation* are the most correlated by linguistic features in the validation set (see Table 3.11) i.e. any modification introduced in the segmented sentence eg. changing the number of unique verbs[20] can lead to change in the model's prediction. For test set, *Temporality* appears to be least influenced by linguistic features, whereas in the validation set, *Action* showed the least impact, indicating inconsistency across datasets. As shown in Table 3.12, lexico-semantic and syntactic feature families account for the majority of significant correlations with model prediction correctness, suggesting that models may be relying on superficial stylistic cues rather than meaningful medical context distinctions. An exception is observed for *Actor* and *Negation*, where discourse-related linguistic features play a dominant role.

Overall, these findings reinforce our initial hypothesis that variation in writing style substantially shapes how models interpret and process clinical narratives which further

[20]See Table 3.13, for total_number_of_unique_verbs we obtain coefficient of -6.60 with a $p$ value of $10^{-10}$.

impacts language model's capability to capture information from rich context such a medical domain.

## Summary

In this study, we examined the influence of clinician writing style variation on the performance of language models for context classification in clinical narratives. Our findings demonstrate that stylistic differences captured through linguistic features can significantly impact how language models interpret medical contexts. These results underscore the importance of accounting for writing style heterogeneity when developing medical NLP systems. Incorporating stylistic normalization or adaptation strategies may improve model robustness, particularly in scenarios with limited annotated clinical data.

**Limitations** While this study highlights the effect of writing style variation on language model performance in the medical domain, several limitations must be acknowledged.

First, the analysis was conducted on a relatively small set of clinical notes drawn from a limited number of healthcare providers, which may not represent the full spectrum of stylistic variation across institutions, specialties, or geographic regions.

Second, our focus was restricted to linguistic features as indicators of stylistic variation. Other influential factors such as differences in clinical expertise, documentation protocols, or institutional practices were not considered.

Third, the investigation approached the problem from a correlation perspective within selected downstream tasks. As such, the findings may not generalize to all potential confounding variables or broader applications.

Finally, our analysis was limited to pretrained language models. Future research should extend this investigation to large language models (LLMs) to determine whether similar stylistic sensitivities persist at larger scales and how they may be mitigated.

| **Action** | Coef. | Std.Err. | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| average_brysbaert_age_of_acquistion_of_words_pe... | -1.332967 | 0.528185 | -2.523674 | 1.388597e-02 | -2.386399 | -0.279535 |
| corrected_adverbs_variation | 5175.404330 | 2283.831956 | 2.266106 | 2.653914e-02 | 620.445122 | 9730.363538 |
| root_adverbs_variation | -5201.900492 | 2296.742110 | -2.264904 | 2.661663e-02 | -9782.608191 | -621.192792 |
| total_number_of_unique_nouns | -13.631678 | 6.105775 | -2.232588 | 2.877686e-02 | -25.809262 | -1.454095 |
| simple_numerals_variation | -12.722929 | 5.826538 | -2.183617 | 3.234429e-02 | -24.343593 | -1.102265 |
| total_kuperman_age_of_acquistion_of_words | 37.321799 | 18.411220 | 2.027123 | 4.645964e-02 | 0.601778 | 74.041819 |
| average_kuperman_age_of_acquistion_of_words_per... | -36.646022 | 18.083067 | -2.026538 | 4.652102e-02 | -72.711562 | -0.580481 |

Table 3.13: Action

| **Actor** | Coef. | Std.Err. | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| total_number_of_unique_verbs | -6.603329e+00 | 8.827628e-01 | -7.480298 | 1.650764e-10 | -8.363944e+00 | -4.842714e+00 |
| total_number_of_named_entities_cardinal | 1.423989e-08 | 2.126493e-09 | 6.696419 | 4.446148e-09 | 9.998731e-09 | 1.848104e-08 |
| average_number_of_adverbs_per_word | 1.320402e+00 | 2.292887e-01 | 5.758689 | 2.077356e-07 | 8.631005e-01 | 1.777704e+00 |
| total_number_of_named_entities_loc | 2.985107e-08 | 5.366359e-09 | 5.562631 | 4.546192e-07 | 1.914821e-08 | 4.055394e-08 |
| average_number_of_named_entities_norp_per_sentence | 3.550854e-08 | 6.615868e-09 | 5.367178 | 9.835470e-07 | 2.231361e-08 | 4.870347e-08 |
| average_number_of_named_entities_ordinal_per_se... | -1.419000e-08 | 2.679636e-09 | -5.295494 | 1.302100e-06 | -1.953436e-08 | -8.845632e-09 |
| total_number_of_named_entities_date | 7.829471e-08 | 1.498465e-08 | 5.224995 | 1.713522e-06 | 4.840878e-08 | 1.081806e-07 |
| simple_verbs_variation | -3.609825e+00 | 7.089304e-01 | -5.091932 | 2.866104e-06 | -5.023743e+00 | -2.195908e+00 |
| total_number_of_unique_adverbs | 7.335495e+00 | 1.457505e+00 | 5.032912 | 3.594641e-06 | 4.428592e+00 | 1.024240e+01 |
| total_number_of_named_entities_money | 6.785820e-08 | 1.353844e-08 | 5.012262 | 3.890109e-06 | 4.085664e-08 | 9.485977e-08 |
| simple_coordinating_conjunctions_variation | 9.094260e+00 | 1.823606e+00 | 4.986966 | 4.284578e-06 | 5.457193e+00 | 1.273133e+01 |
| simple_determiners_variation | -3.627148e+00 | 7.319451e-01 | -4.955492 | 4.830352e-06 | -5.086966e+00 | -2.167329e+00 |
| average_number_of_named_entities_date_per_word | 6.510582e-08 | 1.318313e-08 | 4.938572 | 5.151270e-06 | 3.881290e-08 | 9.139874e-08 |
| simple_adverbs_variation | 9.767410e+00 | 1.989599e+00 | 4.909235 | 5.757781e-06 | 5.799280e+00 | 1.373554e+01 |
| total_number_of_named_entities_norp | -3.300574e-07 | 6.728979e-08 | -4.905014 | 5.850624e-06 | -4.642626e-07 | -1.958521e-07 |
| root_adverbs_variation | -2.730079e+03 | 5.568992e+02 | -4.902286 | 5.911379e-06 | -3.840780e+03 | -1.619379e+03 |
| average_number_of_named_entities_law_per_word | -8.650389e-09 | 1.764684e-09 | -4.901949 | 5.918936e-06 | -1.216994e-08 | -5.130839e-09 |
| corrected_adverbs_variation | 2.712329e+03 | 5.537689e+02 | 4.897944 | 6.009383e-06 | 1.607872e+03 | 3.816786e+03 |
| average_number_of_coordinating_conjunctions_per... | 9.809528e-01 | 2.006043e-01 | 4.889989 | 6.193042e-06 | 5.808601e-01 | 1.381045e+00 |
| average_number_of_syllables_per_word | 2.316189e+02 | 4.836869e+01 | 4.788613 | 9.072603e-06 | 1.351506e+02 | 3.280872e+02 |
| total_number_of_named_entities_ordinal | -2.596594e-08 | 5.438161e-09 | -4.774765 | 9.555796e-06 | -3.681201e-08 | -1.511987e-08 |
| average_number_of_auxiliaries_per_word | 1.251121e+00 | 2.645758e-01 | 4.728781 | 1.134728e-05 | 7.234414e-01 | 1.778801e+00 |
| average_number_of_pronouns_per_word | 1.311673e+00 | 2.800749e-01 | 4.683294 | 1.343990e-05 | 7.530815e-01 | 1.870265e+00 |
| average_number_of_nouns_per_word | 2.257798e+00 | 4.860383e-01 | 4.645308 | 1.547154e-05 | 1.288425e+00 | 3.227171e+00 |
| total_number_of_unique_nouns | 6.716098e+00 | 1.480489e+00 | 4.536405 | 2.309635e-05 | 3.763356e+00 | 9.668840e+00 |
| average_number_of_named_entities_event_per_word | 2.472929e-08 | 5.466195e-09 | 4.524040 | 2.416458e-05 | 1.382730e-08 | 3.563127e-08 |
| total_number_of_unique_determiners | -3.375619e+00 | 7.541151e-01 | -4.476264 | 2.876153e-05 | -4.879654e+00 | -1.871584e+00 |
| root_verbs_variation | 1.868781e+03 | 4.193529e+02 | 4.456344 | 3.091963e-05 | 1.032408e+03 | 2.705154e+03 |
| corrected_verbs_variation | -1.860470e+03 | 4.178374e+02 | -4.452618 | 3.134043e-05 | -2.693820e+03 | -1.027120e+03 |
| average_number_of_named_entities_language_per_word | 4.124240e-08 | 9.411679e-09 | 4.382045 | 4.044268e-05 | 2.247140e-08 | 6.001340e-08 |
| total_number_of_named_entities_language | -4.495325e-08 | 1.068470e-08 | -4.207255 | 7.539254e-05 | -6.626321e-08 | -2.364329e-08 |
| average_number_of_adpositions_per_word | 1.292568e+00 | 3.124311e-01 | 4.137131 | 9.644439e-05 | 6.694440e-01 | 1.915692e+00 |
| root_auxiliaries_variation | 7.020887e+02 | 1.715549e+02 | 4.092502 | 1.126829e-04 | 3.599333e+02 | 1.044244e+03 |
| average_number_of_named_entities_law_per_sentence | -1.571381e-08 | 3.840805e-09 | -4.091279 | 1.131628e-04 | -2.337405e-08 | -8.053563e-09 |
| coleman_liau_index | 3.698407e+00 | 9.062749e-01 | 4.080889 | 1.173224e-04 | 1.890899e+00 | 5.505916e+00 |
| corrected_auxiliaries_variation | -6.962670e+02 | 1.707454e+02 | -4.077808 | 1.185835e-04 | -1.036808e+03 | -3.557260e+02 |
| average_number_of_named_entities_org_per_sentence | 1.349471e-08 | 3.326202e-09 | 4.057091 | 1.274113e-04 | 6.860804e-09 | 2.012861e-08 |
| average_number_of_named_entities_date_per_sentence | 1.978967e-08 | 4.912671e-09 | 4.028292 | 1.407396e-04 | 9.991661e-09 | 2.958769e-08 |
| total_number_of_named_entities_time | 7.983593e-08 | 2.011475e-08 | 3.969023 | 1.725146e-04 | 3.971832e-08 | 1.199535e-07 |
| average_number_of_named_entities_fac_per_sentence | 1.480192e-08 | 3.734692e-09 | 3.963358 | 1.758897e-04 | 7.353312e-09 | 2.225053e-08 |
| average_number_of_named_entities_loc_per_sentence | 1.433150e-08 | 3.616952e-09 | 3.962314 | 1.765185e-04 | 7.117716e-09 | 2.154528e-08 |
| average_number_of_named_entities_language_per_s... | 3.009464e-08 | 7.654272e-09 | 3.931745 | 1.959175e-04 | 1.482868e-08 | 4.536061e-08 |
| average_number_of_spaces_per_word | 1.032595e+00 | 2.675687e-01 | 3.859178 | 2.504993e-04 | 4.989463e-01 | 1.566244e+00 |
| corrected_adjectives_variation | 4.139136e+03 | 1.074507e+03 | 3.852126 | 2.565202e-04 | 1.996099e+03 | 6.282173e+03 |
| root_adjectives_variation | -4.143990e+03 | 1.076025e+03 | -3.851202 | 2.573188e-04 | -6.290054e+03 | -1.997926e+03 |
| smog_index | -7.740051e-01 | 2.018465e-01 | -3.834623 | 2.720674e-04 | -1.176575e+00 | -3.714350e-01 |
| total_number_of_words_more_than_three_syllables | 1.110019e+00 | 2.919974e-01 | 3.801468 | 3.040247e-04 | 5.276483e-01 | 1.692389e+00 |
| average_number_of_verbs_per_word | 1.098956e+00 | 2.953297e-01 | 3.721116 | 3.970383e-04 | 5.099396e-01 | 1.687973e+00 |
| root_coordinating_conjunctions_variation | -5.918521e+03 | 1.630449e+03 | -3.629995 | 5.353012e-04 | -9.170348e+03 | -2.666693e+03 |
| corrected_coordinating_conjunctions_variation | 5.907623e+03 | 1.629720e+03 | 3.624931 | 5.441962e-04 | 2.657249e+03 | 9.157997e+03 |
| average_number_of_proper_nouns_per_word | 1.291358e+00 | 3.613331e-01 | 3.573872 | 6.421061e-04 | 5.707021e-01 | 2.012015e+00 |
| average_number_of_named_entities_quantity_per_s... | -6.742536e-09 | 1.893705e-09 | -3.560499 | 6.703914e-04 | -1.051941e-08 | -2.965659e-09 |
| average_number_of_subordinating_conjunctions_pe... | 4.302424e-01 | 1.223792e-01 | 3.515651 | 7.741333e-04 | 1.861648e-01 | 6.743199e-01 |
| average_number_of_named_entities_cardinal_per_s... | -2.051229e-08 | 5.878661e-09 | -3.489279 | 8.420618e-04 | -3.223691e-08 | -8.787671e-09 |
| average_number_of_named_entities_time_per_word | 2.397853e-08 | 7.173665e-09 | 3.342577 | 1.335283e-03 | 9.671104e-09 | 3.828595e-08 |
| reading_time_for_slow_readers | 2.543545e+01 | 7.808767e+00 | 3.257294 | 1.736213e-03 | 9.861352e+00 | 4.100954e+01 |
| average_number_of_named_entities_ordinal_per_word | -2.398542e-08 | 7.495725e-09 | -3.199880 | 2.067104e-03 | -3.893517e-08 | -9.035665e-09 |

Table 3.14: Actor - Part1

| **Actor** | Coef. | Std.Err. | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| simple_nouns_variation | 4.572277e-01 | 1.447439e-01 | 3.158873 | 2.338662e-03 | 1.685451e-01 | 7.459103e-01 |
| total_number_of_unique_pronouns | -5.933595e+00 | 1.886614e+00 | -3.145102 | 2.437108e-03 | -9.696329e+00 | -2.170861e+00 |
| simple_pronouns_variation | -5.690443e+00 | 1.838407e+00 | -3.095312 | 2.826292e-03 | -9.357029e+00 | -2.023856e+00 |
| average_number_of_numerals_per_word | 9.211165e-01 | 3.016055e-01 | 3.054044 | 3.192013e-03 | 3.195834e-01 | 1.522650e+00 |
| average_number_of_adjectives_per_word | 9.039703e-01 | 2.996039e-01 | 3.017218 | 3.555111e-03 | 3.064291e-01 | 1.501511e+00 |
| average_number_of_named_entities_norp_per_word | 3.013695e-08 | 1.038449e-08 | 2.902112 | 4.952210e-03 | 9.425735e-09 | 5.084816e-08 |
| root_particles_variation | 1.355049e+01 | 4.699383e+00 | 2.883462 | 5.221432e-03 | 4.177868e+00 | 2.292311e+01 |
| corrected_particles_variation | 1.369325e+01 | 4.750674e+00 | 2.882380 | 5.237457e-03 | 4.218327e+00 | 2.316817e+01 |
| root_adpositions_variation | 2.611851e+02 | 9.240457e+01 | 2.826539 | 6.128892e-03 | 7.689001e+01 | 4.454802e+02 |
| total_number_of_named_entities_org | 7.365727e-08 | 2.613092e-08 | 2.818778 | 6.263278e-03 | 2.154080e-08 | 1.257737e-07 |
| corrected_adpositions_variation | -2.592618e+02 | 9.275231e+01 | -2.795206 | 6.688263e-03 | -4.442505e+02 | -7.427314e+01 |
| simple_particles_variation | -1.393290e+01 | 5.031651e+00 | -2.769051 | 7.190673e-03 | -2.396821e+01 | -3.897588e+00 |
| root_determiners_variation | 8.658698e+02 | 3.227626e+02 | 2.682683 | 9.105436e-03 | 2.221401e+02 | 1.509599e+03 |
| total_number_of_unique_numerals | 4.208560e+00 | 1.576313e+00 | 2.669876 | 9.426012e-03 | 1.064703e+00 | 7.352417e+00 |
| corrected_determiners_variation | -8.581236e+02 | 3.219952e+02 | -2.665020 | 9.550229e-03 | -1.500323e+03 | -2.159244e+02 |
| average_number_of_named_entities_gpe_per_word | 1.416567e-08 | 5.513924e-09 | 2.569072 | 1.233065e-02 | 3.168492e-09 | 2.516284e-08 |
| total_number_of_unique_particles | -1.333033e+01 | 5.194925e+00 | -2.566030 | 1.242971e-02 | -2.369128e+01 | -2.969382e+00 |
| average_number_of_named_entities_gpe_per_sentence | -1.223607e-08 | 4.854813e-09 | -2.520399 | 1.400475e-02 | -2.191869e-08 | -2.553447e-09 |
| total_number_of_named_entities_fac | -1.487456e-07 | 6.189584e-08 | -2.403160 | 1.890746e-02 | -2.721930e-07 | -2.529826e-08 |
| total_number_of_named_entities_law | 7.439520e-08 | 3.105679e-08 | 2.395456 | 1.927782e-02 | 1.245437e-08 | 1.363360e-07 |
| simple_adjectives_variation | 3.805237e+00 | 1.590028e+00 | 2.393189 | 1.938805e-02 | 6.340270e-01 | 6.976447e+00 |
| total_number_of_named_entities_quantity | 5.019254e-09 | 2.127752e-09 | 2.358947 | 2.112259e-02 | 7.755860e-10 | 9.262923e-09 |
| average_number_of_named_entities_percent_per_se... | -4.051988e-09 | 1.740027e-09 | -2.328693 | 2.276858e-02 | -7.522361e-09 | -5.816140e-10 |
| average_number_of_named_entities_person_per_sen... | 6.094086e-09 | 2.686051e-09 | 2.268790 | 2.636675e-02 | 7.369257e-10 | 1.145125e-08 |
| root_type_token_ratio | -9.129534e+01 | 4.082464e+01 | -2.236281 | 2.852244e-02 | -1.727175e+02 | -9.873173e+00 |
| root_type_token_ratio_no_lemma | -9.129534e+01 | 4.082464e+01 | -2.236281 | 2.852244e-02 | -1.727175e+02 | -9.873173e+00 |
| corrected_type_token_ratio | 8.965713e+01 | 4.088803e+01 | 2.192748 | 3.165116e-02 | 8.108534e+00 | 1.712057e+02 |
| corrected_type_token_ratio_no_lemma | 8.965713e+01 | 4.088803e+01 | 2.192748 | 3.165116e-02 | 8.108534e+00 | 1.712057e+02 |
| total_number_of_named_entities_gpe | -1.436278e-08 | 6.634177e-09 | -2.164968 | 3.380136e-02 | -2.759423e-08 | -1.131333e-09 |
| simple_auxiliaries_variation | -2.849675e+00 | 1.345381e+00 | -2.118117 | 3.771737e-02 | -5.532954e+00 | -1.663967e-01 |
| average_number_of_named_entities_person_per_word | -2.567186e-08 | 1.213546e-08 | -2.115442 | 3.795238e-02 | -4.987526e-08 | -1.468454e-09 |
| average_number_of_named_entities_percent_per_word | 1.101978e-08 | 5.301360e-09 | 2.078670 | 4.131448e-02 | 4.465514e-10 | 2.159301e-08 |
| average_number_of_particles_per_word | 3.744410e-01 | 1.805495e-01 | 2.073897 | 4.176934e-02 | 1.434636e-02 | 7.345356e-01 |
| simple_adpositions_variation | -9.511057e-01 | 4.714977e-01 | -2.017201 | 4.751080e-02 | -1.891478e+00 | -1.073309e-02 |
| average_number_of_characters_per_word | -2.612976e+02 | 1.305870e+02 | -2.000947 | 4.927703e-02 | -5.217451e+02 | -8.500675e-01 |

Table 3.15: Actor - Part2

| **Certainty** | Coef. | Std.Err. | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| corrected_verbs_variation | -3.212384e+03 | 7.606418e+02 | -4.223254 | 7.125228e-05 | -4.729436e+03 | -1.695331e+03 |
| root_verbs_variation | 3.221559e+03 | 7.634007e+02 | 4.220011 | 7.207333e-05 | 1.699005e+03 | 4.744114e+03 |
| average_number_of_particles_per_word | 1.256040e+00 | 3.286769e-01 | 3.821503 | 2.843090e-04 | 6.005144e-01 | 1.911565e+00 |
| average_number_of_adjectives_per_word | 1.975107e+00 | 5.454067e-01 | 3.621348 | 5.505765e-04 | 8.873279e-01 | 3.062887e+00 |
| total_number_of_unique_nouns | 9.644532e+00 | 2.695120e+00 | 3.578517 | 6.325497e-04 | 4.269285e+00 | 1.501978e+01 |
| average_number_of_nouns_per_word | 3.132926e+00 | 8.847966e-01 | 3.540843 | 7.141205e-04 | 1.368255e+00 | 4.897597e+00 |
| average_number_of_spaces_per_word | 1.704043e+00 | 4.870889e-01 | 3.498423 | 8.178939e-04 | 7.325748e-01 | 2.675511e+00 |
| simple_verbs_variation | -4.506679e+00 | 1.290555e+00 | -3.492047 | 8.346754e-04 | -7.080610e+00 | -1.932748e+00 |
| average_number_of_coordinating_conjunctions_per... | 1.272030e+00 | 3.651852e-01 | 3.483246 | 8.583776e-04 | 5.436909e-01 | 2.000369e+00 |
| average_number_of_proper_nouns_per_word | 2.275784e+00 | 6.577801e-01 | 3.459795 | 9.246828e-04 | 9.638833e-01 | 3.587685e+00 |
| average_number_of_numerals_per_word | 1.845811e+00 | 5.490503e-01 | 3.361825 | 1.257749e-03 | 7.507647e-01 | 2.940857e+00 |
| total_number_of_words_more_than_three_syllables | 1.747913e+00 | 5.315596e-01 | 3.288274 | 1.579013e-03 | 6.877512e-01 | 2.808075e+00 |
| total_number_of_unique_verbs | -5.212617e+00 | 1.607004e+00 | -3.243686 | 1.809805e-03 | -8.417686e+00 | -2.007549e+00 |
| average_number_of_pronouns_per_word | 1.622757e+00 | 5.098556e-01 | 3.182777 | 2.176561e-03 | 6.058818e-01 | 2.639632e+00 |
| average_number_of_verbs_per_word | 1.678832e+00 | 5.376258e-01 | 3.122677 | 2.605744e-03 | 6.065711e-01 | 2.751093e+00 |
| total_number_of_named_entities_gpe | -3.761732e-08 | 1.207703e-08 | -3.114784 | 2.667656e-03 | -6.170419e-08 | -1.353045e-08 |
| average_number_of_adpositions_per_word | 1.745591e+00 | 5.687575e-01 | 3.069130 | 3.053488e-03 | 6.112399e-01 | 2.879942e+00 |
| smog_index | -1.118901e+00 | 3.674465e-01 | -3.045073 | 3.277147e-03 | -1.851750e+00 | -3.860523e-01 |
| simple_determiners_variation | -4.040180e+00 | 1.332452e+00 | -3.032140 | 3.403587e-03 | -6.697671e+00 | -1.382689e+00 |
| total_number_of_unique_numerals | 8.620557e+00 | 2.869560e+00 | 3.004139 | 3.693049e-03 | 2.897399e+00 | 1.434371e+01 |
| coleman_liau_index | 4.932388e+00 | 1.649806e+00 | 2.989677 | 3.851343e-03 | 1.641953e+00 | 8.222822e+00 |
| total_number_of_named_entities_quantity | -1.147253e-08 | 3.873415e-09 | -2.961865 | 4.173613e-03 | -1.919781e-08 | -3.747248e-09 |
| total_number_of_unique_adverbs | 7.468378e+00 | 2.653280e+00 | 2.814772 | 6.333705e-03 | 2.176579e+00 | 1.276018e+01 |
| bilogarithmic_type_token_ratio | -6.978865e-01 | 2.541982e-01 | -2.745443 | 7.673647e-03 | -1.204869e+00 | -1.909043e-01 |
| bilogarithmic_type_token_ratio_no_lemma | -6.978865e-01 | 2.541982e-01 | -2.745443 | 7.673652e-03 | -1.204869e+00 | -1.909042e-01 |
| total_kuperman_age_of_acquistion_of_words | -2.174240e+01 | 8.126806e+00 | -2.675392 | 9.286694e-03 | -3.795080e+01 | -5.533992e+00 |
| average_number_of_named_entities_cardinal_per_word | 1.447710e-08 | 5.514504e-09 | 2.625278 | 1.062418e-02 | 3.478774e-09 | 2.547544e-08 |
| simple_type_token_ratio | 1.541694e+00 | 6.013573e-01 | 2.563691 | 1.250639e-02 | 3.423247e-01 | 2.741063e+00 |
| simple_type_token_ratio_no_lemma | 1.541694e+00 | 6.013573e-01 | 2.563691 | 1.250639e-02 | 3.423247e-01 | 2.741063e+00 |
| average_kuperman_age_of_acquistion_of_words_per... | 2.037079e+01 | 7.981958e+00 | 2.552104 | 1.289251e-02 | 4.451275e+00 | 3.629030e+01 |
| simple_numerals_variation | 6.483988e+00 | 2.571864e+00 | 2.521124 | 1.397837e-02 | 1.354567e+00 | 1.161341e+01 |
| average_number_of_named_entities_product_per_se... | -1.031647e-08 | 4.109783e-09 | -2.510223 | 1.437966e-02 | -1.851318e-08 | -2.119768e-09 |
| total_number_of_unique_determiners | -3.325580e+00 | 1.372811e+00 | -2.422461 | 1.800730e-02 | -6.063565e+00 | -5.875961e-01 |
| average_number_of_adverbs_per_word | 9.997150e-01 | 4.174031e-01 | 2.395083 | 1.929594e-02 | 1.672309e-01 | 1.832199e+00 |
| total_number_of_named_entities_cardinal | -9.126049e-09 | 3.871122e-09 | -2.357468 | 2.120049e-02 | -1.684676e-08 | -1.405338e-09 |
| flesch_kincaid_grade_level | 3.737562e+03 | 1.594155e+03 | 2.344541 | 2.189264e-02 | 5.581205e+02 | 6.917003e+03 |
| flesch_kincaid_reading_ease | 2.999031e+03 | 1.286925e+03 | 2.330386 | 2.267354e-02 | 4.323410e+02 | 5.565722e+03 |
| reading_time_for_fast_readers | -1.356564e+01 | 5.852540e+00 | -2.317907 | 2.338238e-02 | -2.523817e+01 | -1.893119e+00 |
| average_number_of_named_entities_money_per_word | 2.653509e-08 | 1.160470e-08 | 2.286581 | 2.524896e-02 | 3.390247e-09 | 4.967994e-08 |
| root_type_token_ratio_no_lemma | -1.652804e+02 | 7.431822e+01 | -2.223955 | 2.937941e-02 | -3.135034e+02 | -1.705738e+01 |
| root_type_token_ratio | -1.652804e+02 | 7.431822e+01 | -2.223955 | 2.937941e-02 | -3.135034e+02 | -1.705738e+01 |
| total_number_of_unique_particles | 2.098733e+01 | 9.456974e+00 | 2.219244 | 2.971291e-02 | 2.125994e+00 | 3.984867e+01 |
| total_number_of_unique_auxiliaries | -5.758467e+00 | 2.625123e+00 | -2.193599 | 3.158723e-02 | -1.099411e+01 | -5.228236e-01 |
| average_number_of_subordinating_conjunctions_pe... | 4.876051e-01 | 2.227822e-01 | 2.188708 | 3.195621e-02 | 4.328008e-02 | 9.319301e-01 |
| total_number_of_unique_coordinating_conjunctions | 4.736511e+00 | 2.166110e+00 | 2.186644 | 3.211308e-02 | 4.163406e-01 | 9.056682e+00 |
| corrected_type_token_ratio_no_lemma | 1.619719e+02 | 7.443361e+01 | 2.176059 | 3.292807e-02 | 1.351877e+01 | 3.104251e+02 |
| corrected_type_token_ratio | 1.619719e+02 | 7.443361e+01 | 2.176059 | 3.292807e-02 | 1.351877e+01 | 3.104251e+02 |
| average_number_of_proper_nouns_per_sentence | -6.589614e+01 | 3.039285e+01 | -2.168146 | 3.354909e-02 | -1.265128e+02 | -5.279508e+00 |
| total_number_of_named_entities_ordinal | 2.134446e-08 | 9.899768e-09 | 2.156057 | 3.451764e-02 | 1.599997e-09 | 4.108893e-08 |
| total_number_of_unique_proper_nouns | 3.251199e+01 | 1.523059e+01 | 2.134651 | 3.629253e-02 | 2.135541e+00 | 6.288845e+01 |
| total_number_of_proper_nouns | 3.251199e+01 | 1.523059e+01 | 2.134651 | 3.629253e-02 | 2.135541e+00 | 6.288845e+01 |
| root_particles_variation | -1.801303e+01 | 8.554877e+00 | -2.105586 | 3.882938e-02 | -3.507519e+01 | -9.508670e-01 |
| corrected_particles_variation | -1.820879e+01 | 8.648248e+00 | -2.105489 | 3.883811e-02 | -3.545718e+01 | -9.604052e-01 |
| root_determiners_variation | 1.223518e+03 | 5.875652e+02 | 2.082353 | 4.096658e-02 | 5.165614e+01 | 2.395380e+03 |
| average_number_of_named_entities_law_per_word | 6.668953e-09 | 3.212475e-09 | 2.075955 | 4.157272e-02 | 2.618729e-10 | 1.307603e-08 |
| corrected_determiners_variation | -1.214813e+03 | 5.861683e+02 | -2.072465 | 4.190660e-02 | -2.383889e+03 | -4.573735e+01 |
| average_number_of_verbs_per_sentence | -5.335525e+01 | 2.651491e+01 | -2.012273 | 4.804043e-02 | -1.062376e+02 | -4.729229e-01 |
| average_number_of_auxiliaries_per_sentence | -6.196236e+01 | 3.080374e+01 | -2.011520 | 4.812178e-02 | -1.233985e+02 | -5.262309e-01 |
| average_number_of_adpositions_per_sentence | -8.579195e+01 | 4.280723e+01 | -2.004146 | 4.892499e-02 | -1.711683e+02 | -4.156157e-01 |
| automated_readability_index | -6.747181e+02 | 3.371693e+02 | -2.001125 | 4.925731e-02 | -1.347181e+03 | -2.255082e+00 |

Table 3.16: Certainty

| **Negation** | Coef. | Std.Err. | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| simple_adpositions_variation | -1.719098e+00 | 6.155434e-01 | -2.792814 | 6.732834e-03 | -2.946761e+00 | -4.914355e-01 |
| total_number_of_named_entities_product | 4.244266e-08 | 1.733488e-08 | 2.448397 | 1.685814e-02 | 7.869341e-09 | 7.701598e-08 |
| smog_index | -6.244617e-01 | 2.635119e-01 | -2.369767 | 2.056014e-02 | -1.150020e+00 | -9.890378e-02 |
| average_number_of_named_entities_loc_per_word | 5.714464e-08 | 2.488083e-08 | 2.296733 | 2.463010e-02 | 7.521382e-09 | 1.067679e-07 |
| average_number_of_named_entities_time_per_sentence | -1.229495e-08 | 5.436076e-09 | -2.261733 | 2.682211e-02 | -2.313686e-08 | -1.453039e-09 |
| total_number_of_named_entities_percent | 9.312301e-08 | 4.136844e-08 | 2.251064 | 2.752364e-02 | 1.061625e-08 | 1.756298e-07 |
| total_number_of_named_entities_law | 9.120490e-08 | 4.054485e-08 | 2.249482 | 2.762905e-02 | 1.034075e-08 | 1.720690e-07 |
| average_number_of_named_entities_cardinal_per_s... | -1.719405e-08 | 7.674631e-09 | -2.240375 | 2.824268e-02 | -3.250062e-08 | -1.887482e-09 |
| average_number_of_named_entities_gpe_per_sentence | -1.393150e-08 | 6.337991e-09 | -2.198094 | 3.125139e-02 | -2.657222e-08 | -1.290775e-09 |
| average_number_of_named_entities_org_per_word | 4.446542e-08 | 2.024354e-08 | 2.196524 | 3.136830e-02 | 4.090961e-09 | 8.483988e-08 |
| average_number_of_named_entities_money_per_sent... | 2.024129e-08 | 9.350754e-09 | 2.164669 | 3.382518e-02 | 1.591797e-09 | 3.889078e-08 |
| total_number_of_named_entities_fac | -1.743072e-07 | 8.080543e-08 | -2.157123 | 3.443124e-02 | -3.354686e-07 | -1.314590e-08 |
| total_number_of_named_entities_event | 6.852808e-08 | 3.183012e-08 | 2.152932 | 3.477191e-02 | 5.044914e-09 | 1.320112e-07 |
| average_number_of_named_entities_art_per_sentence | 2.453337e-08 | 1.147241e-08 | 2.138468 | 3.597032e-02 | 1.652384e-09 | 4.741436e-08 |
| average_number_of_named_entities_loc_per_sentence | 1.000902e-08 | 4.721955e-09 | 2.119678 | 3.758082e-02 | 5.913809e-10 | 1.942667e-08 |
| average_number_of_named_entities_gpe_per_word | 1.519277e-08 | 7.198464e-09 | 2.110557 | 3.838489e-02 | 8.358857e-10 | 2.954965e-08 |
| average_number_of_named_entities_art_per_word | -5.169465e-08 | 2.462123e-08 | -2.099597 | 3.937085e-02 | -1.008001e-07 | -2.589165e-09 |
| simple_coordinating_conjunctions_variation | 4.958396e+00 | 2.380729e+00 | 2.082721 | 4.093188e-02 | 2.101811e-01 | 9.706611e+00 |
| average_number_of_named_entities_percent_per_word | 1.438223e-08 | 6.920960e-09 | 2.078069 | 4.137159e-02 | 5.788102e-10 | 2.818565e-08 |
| average_number_of_named_entities_fac_per_word | -3.128243e-08 | 1.506596e-08 | -2.076365 | 4.153365e-02 | -6.133055e-08 | -1.234316e-09 |
| root_coordinating_conjunctions_variation | -4.413945e+03 | 2.128561e+03 | -2.073675 | 4.179054e-02 | -8.659227e+03 | -1.686630e+02 |
| corrected_coordinating_conjunctions_variation | 4.409099e+03 | 2.127610e+03 | 2.072325 | 4.192006e-02 | 1.657140e+02 | 8.652483e+03 |
| average_number_of_named_entities_event_per_sent... | 1.840485e-08 | 8.899373e-09 | 2.068107 | 4.232679e-02 | 6.556136e-10 | 3.615409e-08 |
| average_number_of_named_entities_person_per_word | -3.237656e-08 | 1.584292e-08 | -2.043598 | 4.475826e-02 | -6.397426e-08 | -7.788546e-10 |
| average_number_of_named_entities_language_per_word | 2.490433e-08 | 1.228701e-08 | 2.026883 | 4.648480e-02 | 3.986602e-10 | 4.941000e-08 |
| total_number_of_named_entities_person | 8.564630e-08 | 4.238720e-08 | 2.020570 | 4.715160e-02 | 1.107697e-09 | 1.701849e-07 |

Table 3.17: Negation

| **Temporality** | Coef. | Std.Err. | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| total_number_of_unique_punctuations | -8.488106e+00 | 1.458012e+00 | -5.821699 | 1.612211e-07 | -1.139602e+01 | -5.580193e+00 |
| simple_type_token_ratio_no_lemma | 4.004287e+00 | 6.977944e-01 | 5.738491 | 2.252798e-07 | 2.612580e+00 | 5.395994e+00 |
| simple_type_token_ratio | 4.004287e+00 | 6.977943e-01 | 5.738491 | 2.252798e-07 | 2.612580e+00 | 5.395994e+00 |
| simple_punctuations_variation | -5.202325e+00 | 9.263226e-01 | -5.616105 | 3.674957e-07 | -7.049817e+00 | -3.354833e+00 |
| total_number_of_unique_nouns | 1.714003e+01 | 3.127325e+00 | 5.480731 | 6.289046e-07 | 1.090277e+01 | 2.337728e+01 |
| total_number_of_unique_particles | 5.991561e+01 | 1.097355e+01 | 5.460003 | 6.825611e-07 | 3.802956e+01 | 8.180166e+01 |
| simple_auxiliaries_variation | -1.550611e+01 | 2.841929e+00 | -5.456191 | 6.929087e-07 | -2.117416e+01 | -9.838059e+00 |
| corrected_particles_variation | -5.234579e+01 | 1.003513e+01 | -5.216255 | 1.772690e-06 | -7.236023e+01 | -3.233136e+01 |
| root_particles_variation | -5.177689e+01 | 9.926785e+00 | -5.215876 | 1.775294e-06 | -7.157523e+01 | -3.197854e+01 |
| average_number_of_punctuations_per_word | -3.707239e+00 | 7.182851e-01 | -5.161235 | 2.193892e-06 | -5.139813e+00 | -2.274664e+00 |
| simple_particles_variation | 5.464536e+01 | 1.062865e+01 | 5.141324 | 2.369345e-06 | 3.344717e+01 | 7.584354e+01 |
| total_number_of_unique_auxiliaries | -1.535464e+01 | 3.046103e+00 | -5.040749 | 3.488348e-06 | -2.142990e+01 | -9.279380e+00 |
| coleman_liau_index | 9.580976e+00 | 1.914378e+00 | 5.004745 | 4.003460e-06 | 5.762868e+00 | 1.339908e+01 |
| average_number_of_adpositions_per_word | 3.269788e+00 | 6.599667e-01 | 4.954474 | 4.849078e-06 | 1.953526e+00 | 4.586050e+00 |
| total_number_of_words_more_than_three_syllables | 2.960606e+00 | 6.168035e-01 | 4.799918 | 8.695859e-06 | 1.730431e+00 | 4.190782e+00 |
| simple_verbs_variation | -7.145826e+00 | 1.497516e+00 | -4.771787 | 9.662941e-06 | -1.013253e+01 | -4.159125e+00 |
| average_number_of_named_entities_quantity_per_word | -4.121338e-08 | 8.651680e-09 | -4.763627 | 9.962572e-06 | -5.846861e-08 | -2.395814e-08 |
| average_number_of_interjections_per_word | 1.792648e+00 | 3.835171e-01 | 4.674232 | 1.389958e-05 | 1.027747e+00 | 2.557549e+00 |
| corrected_nouns_variation | -9.959689e+02 | 2.159900e+02 | -4.611182 | 1.754958e-05 | -1.426747e+03 | -5.651905e+02 |
| root_nouns_variation | 9.975494e+02 | 2.166140e+02 | 4.605194 | 1.794119e-05 | 5.655264e+02 | 1.429572e+03 |
| simple_proper_nouns_variation | 9.242736e+00 | 2.047195e+00 | 4.514829 | 2.499143e-05 | 5.159734e+00 | 1.332574e+01 |
| average_number_of_named_entities_art_per_word | 1.797830e-07 | 3.983808e-08 | 4.512842 | 2.517333e-05 | 1.003284e-07 | 2.592375e-07 |
| average_number_of_named_entities_percent_per_word | -5.033151e-08 | 1.119838e-08 | -4.494536 | 2.691110e-05 | -7.266598e-08 | -2.799705e-08 |
| total_number_of_unique_adpositions | 1.096727e+01 | 2.450401e+00 | 4.475704 | 2.882018e-05 | 6.080100e+00 | 1.585444e+01 |
| total_number_of_words_more_than_two_syllables | 4.532242e+00 | 1.019734e+00 | 4.444535 | 3.227238e-05 | 2.498447e+00 | 6.566036e+00 |
| root_proper_nouns_variation | -3.887525e+03 | 8.832845e+02 | -4.401215 | 3.774369e-05 | -5.649180e+03 | -2.125870e+03 |
| corrected_proper_nouns_variation | 3.870173e+03 | 8.797788e+02 | 4.399030 | 3.804234e-05 | 2.115510e+03 | 5.624837e+03 |
| root_verbs_variation | 3.864364e+03 | 8.858240e+02 | 4.362452 | 4.339410e-05 | 2.097644e+03 | 5.631085e+03 |
| corrected_verbs_variation | -3.847501e+03 | 8.826227e+02 | -4.359168 | 4.390870e-05 | -5.607836e+03 | -2.087165e+03 |
| average_number_of_named_entities_fac_per_word | 1.061814e-07 | 2.437730e-08 | 4.355750 | 4.445068e-05 | 5.756244e-08 | 1.548004e-07 |
| root_punctuations_variation | 7.581011e+02 | 1.749593e+02 | 4.333015 | 4.822404e-05 | 4.091559e+02 | 1.107046e+03 |
| total_number_of_named_entities_event | -2.224247e-07 | 5.150234e-08 | -4.318730 | 5.075161e-05 | -3.251429e-07 | -1.197066e-07 |
| corrected_punctuations_variation | -7.458231e+02 | 1.732774e+02 | -4.304215 | 5.345123e-05 | -1.091414e+03 | -4.002322e+02 |
| total_number_of_named_entities_product | -1.185646e-07 | 2.804849e-08 | -4.227129 | 7.028313e-05 | -1.745055e-07 | -6.262363e-08 |
| average_number_of_named_entities_event_per_sent... | -6.009121e-08 | 1.439952e-08 | -4.173139 | 8.501091e-05 | -8.881016e-08 | -3.137227e-08 |
| average_subtlex_us_zipf_of_words_per_word | -3.918426e+00 | 9.524894e-01 | -4.113878 | 1.046008e-04 | -5.818106e+00 | -2.018745e+00 |
| average_number_of_numerals_per_word | 2.587804e+00 | 6.370992e-01 | 4.061854 | 1.253275e-04 | 1.317149e+00 | 3.858458e+00 |
| average_number_of_auxiliaries_per_word | 2.262518e+00 | 5.588793e-01 | 4.048313 | 1.313395e-04 | 1.147869e+00 | 3.377168e+00 |
| gunning_fog_index | -2.912632e+00 | 7.260163e-01 | -4.011800 | 1.489660e-04 | -4.360626e+00 | -1.464638e+00 |
| simple_nouns_variation | -1.215635e+00 | 3.057511e-01 | -3.975896 | 1.685040e-04 | -1.825436e+00 | -6.058333e-01 |
| total_kuperman_age_of_acquistion_of_words | -3.737562e+01 | 9.430067e+00 | -3.963452 | 1.758329e-04 | -5.618330e+01 | -1.856795e+01 |
| total_number_of_unique_spaces | 5.749141e+01 | 1.456223e+01 | 3.947981 | 1.853725e-04 | 2.844796e+01 | 8.653487e+01 |
| total_number_of_named_entities_percent | -2.633546e-07 | 6.693571e-08 | -3.934441 | 1.941272e-04 | -3.968537e-07 | -1.298556e-07 |
| average_number_of_named_entities_org_per_word | -1.288588e-07 | 3.275481e-08 | -3.934041 | 1.943917e-04 | -1.941862e-07 | -6.353138e-08 |
| average_number_of_named_entities_loc_per_word | -1.580689e-07 | 4.025814e-08 | -3.926385 | 1.995235e-04 | -2.383613e-07 | -7.777661e-08 |
| root_adjectives_variation | -8.902584e+03 | 2.272952e+03 | -3.916750 | 2.061655e-04 | -1.343584e+04 | -4.369325e+03 |
| corrected_adjectives_variation | 8.889159e+03 | 2.269745e+03 | 3.916369 | 2.064328e-04 | 4.362295e+03 | 1.341602e+04 |
| average_number_of_nouns_per_word | 4.013621e+00 | 1.026688e+00 | 3.909291 | 2.114528e-04 | 1.965957e+00 | 6.061285e+00 |
| average_number_of_named_entities_time_per_sentence | 3.425673e-08 | 8.795779e-09 | 3.894679 | 2.221910e-04 | 1.671410e-08 | 5.179936e-08 |
| corrected_spaces_variation | -5.365895e+01 | 1.386591e+01 | -3.869848 | 2.416467e-04 | -8.131363e+01 | -2.600427e+01 |
| simple_spaces_variation | 5.451684e+01 | 1.428988e+01 | 3.815067 | 2.905048e-04 | 2.601658e+01 | 8.301710e+01 |
| average_kuperman_age_of_acquistion_of_words_per... | 3.524573e+01 | 9.261990e+00 | 3.805416 | 3.000387e-04 | 1.677327e+01 | 5.371818e+01 |
| reading_time_for_fast_readers | -2.574417e+01 | 6.791087e+00 | -3.790877 | 3.149690e-04 | -3.928857e+01 | -1.219978e+01 |
| root_spaces_variation | -4.460722e+01 | 1.180521e+01 | -3.778605 | 3.281203e-04 | -6.815197e+01 | -2.106248e+01 |
| average_number_of_named_entities_money_per_sent... | -5.695573e-08 | 1.512988e-08 | -3.764454 | 3.439378e-04 | -8.713132e-08 | -2.678014e-08 |
| average_number_of_named_entities_person_per_word | 9.622665e-08 | 2.563444e-08 | 3.753803 | 3.563216e-04 | 4.510036e-08 | 1.473529e-07 |
| total_brysbaert_age_of_acquistion_of_words | 2.551302e+01 | 6.819914e+00 | 3.740959 | 3.718210e-04 | 1.191113e+01 | 3.911491e+01 |
| total_number_of_named_entities_law | -2.418963e-07 | 6.560311e-08 | -3.687269 | 4.438607e-04 | -3.727376e-07 | -1.110550e-07 |
| average_number_of_named_entities_art_per_sentence | -6.825043e-08 | 1.856279e-08 | -3.676733 | 4.594789e-04 | -1.052727e-07 | -3.122811e-08 |
| total_number_of_unique_words | -2.855957e+01 | 7.842706e+00 | -3.641545 | 5.155262e-04 | -4.420135e+01 | -1.291778e+01 |
| total_number_of_named_entities_fac | 4.710148e-07 | 1.307463e-07 | 3.602511 | 5.852995e-04 | 2.102496e-07 | 7.317800e-07 |
| average_number_of_named_entities_percent_per_se... | -1.321096e-08 | 3.675562e-09 | -3.594270 | 6.011350e-04 | -2.054164e-08 | -5.880286e-09 |

Table 3.18: Temporality-Part1

| **Temporality** | Coef. | Std.Err. | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| total_number_of_named_entities_art | 1.704759e-07 | 4.760417e-08 | 3.581112 | 6.272691e-04 | 7.553235e-08 | 2.654194e-07 |
| average_brysbaert_age_of_acquistion_of_words_pe... | -2.492547e+01 | 7.055126e+00 | -3.532959 | 7.324102e-04 | -3.899647e+01 | -1.085447e+01 |
| average_number_of_named_entities_product_per_word | 9.089588e-08 | 2.590158e-08 | 3.509279 | 7.900521e-04 | 3.923681e-08 | 1.425550e-07 |
| total_number_of_sentences | 2.287950e+01 | 6.519742e+00 | 3.509265 | 7.900892e-04 | 9.876285e+00 | 3.588272e+01 |
| average_number_of_verbs_per_word | 2.158172e+00 | 6.238425e-01 | 3.459483 | 9.255977e-04 | 9.139578e-01 | 3.402387e+00 |
| average_number_of_named_entities_cardinal_per_s... | 4.226246e-08 | 1.241785e-08 | 3.403365 | 1.104648e-03 | 1.749585e-08 | 6.702907e-08 |
| simple_adpositions_variation | 3.327248e+00 | 9.959727e-01 | 3.340702 | 1.343072e-03 | 1.340843e+00 | 5.313653e+00 |
| average_number_of_named_entities_person_per_sen... | 1.894099e-08 | 5.673906e-09 | 3.338263 | 1.353269e-03 | 7.624741e-09 | 3.025724e-08 |
| average_number_of_particles_per_word | 1.264886e+00 | 3.813854e-01 | 3.316555 | 1.447285e-03 | 5.042365e-01 | 2.025535e+00 |
| simple_pronouns_variation | -1.276292e+01 | 3.883376e+00 | -3.286553 | 1.587379e-03 | -2.050807e+01 | -5.017773e+00 |
| average_kuperman_age_of_acquistion_of_words_per... | 9.428420e-01 | 2.882325e-01 | 3.271116 | 1.664341e-03 | 3.679803e-01 | 1.517704e+00 |
| average_number_of_named_entities_language_per_word | -6.366250e-08 | 1.988085e-08 | -3.202202 | 2.052643e-03 | -1.033136e-07 | -2.401140e-08 |
| bilogarithmic_type_token_ratio | -9.381622e-01 | 2.949628e-01 | -3.180611 | 2.190802e-03 | -1.526447e+00 | -3.498773e-01 |
| bilogarithmic_type_token_ratio_no_lemma | -9.381621e-01 | 2.949628e-01 | -3.180611 | 2.190803e-03 | -1.526447e+00 | -3.498773e-01 |
| average_number_of_named_entities_gpe_per_sentence | 3.259444e-08 | 1.025511e-08 | 3.178361 | 2.205693e-03 | 1.214127e-08 | 5.304762e-08 |
| average_number_of_named_entities_norp_per_word | 6.964400e-08 | 2.193577e-08 | 3.174905 | 2.228745e-03 | 2.589448e-08 | 1.133935e-07 |
| average_number_of_characters_per_sentence | -1.453636e+02 | 4.604424e+01 | -3.157042 | 2.351536e-03 | -2.371960e+02 | -5.353126e+01 |
| total_number_of_named_entities | -2.994569e+01 | 9.508760e+00 | -3.149274 | 2.406884e-03 | -4.891031e+01 | -1.098107e+01 |
| smog_index | -1.334089e+00 | 4.263723e-01 | -3.128930 | 2.557660e-03 | -2.184462e+00 | -4.837163e-01 |
| total_number_of_characters | 1.355410e+02 | 4.332840e+01 | 3.128224 | 2.563044e-03 | 4.912519e+01 | 2.219567e+02 |
| average_number_of_stop_words_per_word | 2.581518e+00 | 8.283695e-01 | 3.116384 | 2.654992e-03 | 9.293867e-01 | 4.233648e+00 |
| total_number_of_named_entities_gpe | -4.354429e-08 | 1.401377e-08 | -3.107251 | 2.728013e-03 | -7.149386e-08 | -1.559471e-08 |
| average_number_of_named_entities_per_sentence | 2.983344e+01 | 9.603292e+00 | 3.106585 | 2.733412e-03 | 1.068028e+01 | 4.898660e+01 |
| root_coordinating_conjunctions_variation | 1.058921e+04 | 3.444094e+03 | 3.074600 | 3.004668e-03 | 3.720180e+03 | 1.745824e+04 |
| average_number_of_named_entities_quantity_per_s... | -1.229481e-08 | 4.000186e-09 | -3.073559 | 3.013899e-03 | -2.027293e-08 | -4.316689e-09 |
| corrected_coordinating_conjunctions_variation | -1.057527e+04 | 3.442554e+03 | -3.071925 | 3.028455e-03 | -1.744123e+04 | -3.709309e+03 |
| simple_coordinating_conjunctions_variation | -1.181247e+01 | 3.852111e+00 | -3.066492 | 3.077308e-03 | -1.949526e+01 | -4.129673e+00 |
| corrected_adpositions_variation | -5.967113e+02 | 1.959262e+02 | -3.045592 | 3.272164e-03 | -9.874738e+02 | -2.059487e+02 |
| average_number_of_spaces_per_word | 1.706844e+00 | 5.652012e-01 | 3.019887 | 3.527553e-03 | 5.795856e-01 | 2.834102e+00 |
| root_adpositions_variation | 5.875629e+02 | 1.951917e+02 | 3.010184 | 3.628681e-03 | 1.982654e+02 | 9.768603e+02 |
| root_pronouns_variation | 1.358380e+03 | 4.603614e+02 | 2.950682 | 4.310101e-03 | 4.402180e+02 | 2.276542e+03 |
| corrected_pronouns_variation | -1.332799e+03 | 4.538840e+02 | -2.936432 | 4.490009e-03 | -2.238043e+03 | -4.275563e+02 |
| average_number_of_named_entities_time_per_word | -4.445738e-08 | 1.515336e-08 | -2.933830 | 4.523594e-03 | -7.467980e-08 | -1.423496e-08 |
| average_number_of_stop_words_per_sentence | -1.660031e+02 | 5.700572e+01 | -2.912043 | 4.814133e-03 | -2.796974e+02 | -5.230877e+01 |
| average_number_of_syllables_per_word | -2.972141e+02 | 1.021721e+02 | -2.908957 | 4.856657e-03 | -5.009898e+02 | -9.343833e+01 |
| uber_type_token_ratio | -3.692130e-01 | 1.272081e-01 | -2.902434 | 4.947675e-03 | -6.229215e-01 | -1.155045e-01 |
| uber_type_token_ratio_no_lemma | -3.692127e-01 | 1.272080e-01 | -2.902432 | 4.947696e-03 | -6.229211e-01 | -1.155043e-01 |
| average_number_of_subordinating_conjunctions_pe... | 7.495760e-01 | 2.585088e-01 | 2.899615 | 4.987489e-03 | 2.339965e-01 | 1.265155e+00 |
| average_number_of_named_entities_per_word | 8.542054e-01 | 2.951073e-01 | 2.894558 | 5.059662e-03 | 2.656324e-01 | 1.442778e+00 |
| average_number_of_interjections_per_sentence | -6.440576e+00 | 2.232746e+00 | -2.884599 | 5.204637e-03 | -1.089365e+01 | -1.987505e+00 |
| total_number_of_puntuations | 3.316925e+01 | 1.154945e+01 | 2.871933 | 5.394545e-03 | 1.013460e+01 | 5.620390e+01 |
| total_number_of_punctuations | 3.316925e+01 | 1.154945e+01 | 2.871933 | 5.394546e-03 | 1.013460e+01 | 5.620390e+01 |
| average_number_of_subordinating_conjunctions_pe... | -1.982583e+01 | 6.934693e+00 | -2.858934 | 5.596079e-03 | -3.365664e+01 | -5.995022e+00 |
| simple_adjectives_variation | 9.458067e+00 | 3.358710e+00 | 2.815982 | 6.312358e-03 | 2.759332e+00 | 1.615680e+01 |
| corrected_numerals_variation | 2.361105e+03 | 8.449339e+02 | 2.794426 | 6.702768e-03 | 6.759375e+02 | 4.046273e+03 |
| root_numerals_variation | -2.369576e+03 | 8.488371e+02 | -2.791555 | 6.756402e-03 | -4.062528e+03 | -6.766232e+02 |
| average_number_of_named_entities_gpe_per_word | -3.240184e-08 | 1.164739e-08 | -2.781897 | 6.939742e-03 | -5.563182e-08 | -9.171852e-09 |
| total_number_of_stop_words | 1.489059e+02 | 5.454004e+01 | 2.730213 | 8.000787e-03 | 4.012924e+01 | 2.576826e+02 |
| total_number_of_named_entities_person | -1.822952e-07 | 6.858410e-08 | -2.657981 | 9.732926e-03 | -3.190819e-07 | -4.550856e-08 |
| average_number_of_punctuations_per_sentence | -5.759106e+01 | 2.174557e+01 | -2.648404 | 9.986602e-03 | -1.009612e+02 | -1.422089e+01 |
| total_number_of_named_entities_quantity | -1.189930e-08 | 4.494578e-09 | -2.647478 | 1.001144e-02 | -2.086345e-08 | -2.935145e-09 |
| simple_adverbs_variation | -1.106562e+01 | 4.202748e+00 | -2.632950 | 1.040869e-02 | -1.944774e+01 | -2.683507e+00 |
| average_number_of_named_entities_date_per_sentence | -2.714601e-08 | 1.037733e-08 | -2.615896 | 1.089321e-02 | -4.784293e-08 | -6.449081e-09 |
| average_number_of_named_entities_loc_per_sentence | -1.948994e-08 | 7.640303e-09 | -2.550938 | 1.293195e-02 | -3.472805e-08 | -4.251837e-09 |
| total_number_of_unique_adjectives | 5.061019e+01 | 1.986889e+01 | 2.547207 | 1.305892e-02 | 1.098293e+01 | 9.023744e+01 |
| total_number_of_spaces | -3.129192e+01 | 1.242049e+01 | -2.519378 | 1.404195e-02 | -5.606381e+01 | -6.520029e+00 |
| average_number_of_named_entities_language_per_s... | -4.039236e-08 | 1.616857e-08 | -2.498202 | 1.483419e-02 | -7.263956e-08 | -8.145154e-09 |
| root_adverbs_variation | 2.935009e+03 | 1.176371e+03 | 2.494968 | 1.495863e-02 | 5.888104e+02 | 5.281208e+03 |
| total_number_of_unique_pronouns | -9.932665e+00 | 3.985208e+00 | -2.492383 | 1.505879e-02 | -1.788091e+01 | -1.984419e+00 |
| corrected_adverbs_variation | -2.913601e+03 | 1.169759e+03 | -2.490770 | 1.512159e-02 | -5.246611e+03 | -5.805901e+02 |

Table 3.19: Temporality-Part2

| **Temporality** | Coef. | Std.Err. | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| total_number_of_adverbs | -4.744528e+01 | 1.920919e+01 | -2.469926 | 1.595468e-02 | -8.575681e+01 | -9.133750e+00 |
| total_number_of_named_entities_loc | -2.761345e-08 | 1.133568e-08 | -2.435977 | 1.739997e-02 | -5.022175e-08 | -5.005156e-09 |
| total_number_of_unique_determiners | 3.816127e+00 | 1.592962e+00 | 2.395617 | 1.927005e-02 | 6.390641e-01 | 6.993190e+00 |
| total_number_of_subordinating_conjunctions | 1.573261e+01 | 6.600784e+00 | 2.383445 | 1.986821e-02 | 2.567758e+00 | 2.889746e+01 |
| total_subtlex_us_zipf_of_words | 5.537261e+01 | 2.323442e+01 | 2.383214 | 1.987971e-02 | 9.033015e+00 | 1.017122e+02 |
| total_number_of_syllables | -1.370393e+02 | 5.794875e+01 | -2.364836 | 2.081476e-02 | -2.526144e+02 | -2.146416e+01 |
| average_number_of_named_entities_ordinal_per_word | 3.711067e-08 | 1.583367e-08 | 2.343783 | 2.193388e-02 | 5.531423e-09 | 6.868993e-08 |
| automated_readability_index | -9.018830e+02 | 3.912397e+02 | -2.305193 | 2.412469e-02 | -1.682186e+03 | -1.215800e+02 |
| total_number_of_named_entities_date | -7.290189e-08 | 3.165296e-08 | -2.303162 | 2.424519e-02 | -1.360317e-07 | -9.772050e-09 |
| average_number_of_characters_per_word | 6.271808e+02 | 2.758466e+02 | 2.273658 | 2.605667e-02 | 7.702202e+01 | 1.177340e+03 |
| average_number_of_adjectives_per_word | 1.435990e+00 | 6.328712e-01 | 2.269008 | 2.635280e-02 | 1.737679e-01 | 2.698212e+00 |
| total_number_of_named_entities_language | -5.113484e-08 | 2.256992e-08 | -2.265619 | 2.657052e-02 | -9.614914e-08 | -6.120547e-09 |
| total_number_of_pronouns | -1.088058e+02 | 4.855724e+01 | -2.240773 | 2.821561e-02 | -2.056501e+02 | -1.196139e+01 |
| total_number_of_determiners | -1.064012e+02 | 4.761202e+01 | -2.234755 | 2.862733e-02 | -2.013604e+02 | -1.144201e+01 |
| total_number_of_named_entities_time | -9.358431e-08 | 4.248959e-08 | -2.202523 | 3.092355e-02 | -1.783271e-07 | -8.841489e-09 |
| average_number_of_named_entities_event_per_word | 2.519426e-08 | 1.154657e-08 | 2.181970 | 3.247077e-02 | 2.165356e-09 | 4.822316e-08 |
| average_number_of_adverbs_per_word | 1.048742e+00 | 4.843402e-01 | 2.165301 | 3.377489e-02 | 8.275605e-02 | 2.014728e+00 |
| average_number_of_pronouns_per_sentence | 9.992538e+01 | 4.657076e+01 | 2.145668 | 3.536939e-02 | 7.042922e+00 | 1.928078e+02 |
| average_number_of_syllables_per_sentence | 1.298597e+02 | 6.068938e+01 | 2.139743 | 3.586327e-02 | 8.818525e+00 | 2.509008e+02 |
| average_number_of_determiners_per_sentence | 1.010542e+02 | 4.728980e+01 | 2.136913 | 3.610127e-02 | 6.737672e+00 | 1.953707e+02 |
| simple_determiners_variation | 3.261927e+00 | 1.546131e+00 | 2.109735 | 3.845809e-02 | 1.782658e-01 | 6.345588e+00 |
| average_number_of_coordinating_conjunctions_per... | 5.408573e+01 | 2.591138e+01 | 2.087335 | 4.049987e-02 | 2.407114e+00 | 1.057643e+02 |
| total_number_of_nouns | -1.968856e+02 | 9.514501e+01 | -2.069322 | 4.220928e-02 | -3.866464e+02 | -7.124892e+00 |
| total_number_of_adjectives | -1.159769e+02 | 5.607170e+01 | -2.068368 | 4.230152e-02 | -2.278084e+02 | -4.145414e+00 |
| total_number_of_coordinating_conjunctions | -5.389970e+01 | 2.619666e+01 | -2.057503 | 4.336443e-02 | -1.061473e+02 | -1.652108e+00 |
| average_number_of_pronouns_per_word | 1.197110e+00 | 5.916190e-01 | 2.023447 | 4.684669e-02 | 1.716281e-02 | 2.377056e+00 |

Table 3.20: Temporality-Part3

## 3.4 Document-level Problems

Many medical NLP tasks require reasoning beyond the sentence or paragraph level, operating instead on entire documents. These documents (for eg. fig. 3.17) may include full clinical notes, discharge summaries, radiology reports, or even scientific abstracts and full-text articles. Unlike token-level tasks such as named entity recognition or relation extraction, document-level tasks introduces a distinct set of challenges that are amplified in the medical domain.

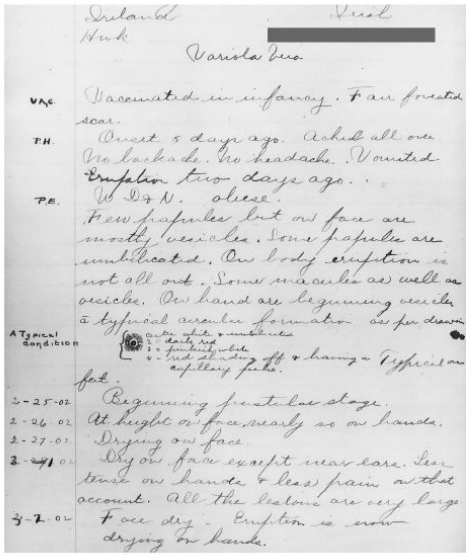

Varicella Duo.
Vaccinated in infancy. Few faint scar.
Onset 5 days ago. Ached all over. No backache. No headache. Vomited Eruption two days ago.
W. D. & N. obese.
Few papules but on face are mostly vesicles. Some papules are umbilicated. On body eruption is not all out. Some macules as well as vesicles. On hand are beginning vesicles of typical areolar formation as per drawing.
Typical condition
2-25-02 Beginning pustular stage.
2-26-02 At height on face, nearly so on hands.
2-27-02 Drying on face.
Dry on face except near ears. Tense on hands & less pain on that account. All the lesions are very large.
3-7-02 Face dry. Eruption is now drying on hands.

Figure 3.17: Clinical Note sample from clinical records of the Southampton Street smallpox hospital (1902).

First, the length of medical documents often exceeds the input limits of standard transformer-based models, requiring specialized handling through truncation, hierarchical modeling, or sparse attention mechanisms. Second, medical documents frequently contain both highly relevant and redundant or boilerplate content. Distinguishing clinically meaningful information from routine documentation, such as templated phrases or repeated sections, is non-trivial and affects both performance and interpretability. Moreover, clinical reasoning is inherently distributed across the document. Important inferences often depend on connecting pieces of information that are separated by many sentences or even sections. This includes resolving coreference across multiple mentions, understanding temporal and causal relationships, and aggregating dispersed evidence for decision-making. Such complexity necessitates models that can maintain coherence, contextual memory, and global understanding over long inputs.

As document-level tasks grow in importance—ranging from patient summarization and cohort selection to literature-based discovery and clinical question answering—it becomes increasingly important to understand the unique obstacles posed at this scale. In the following case studies, we explore a following research question :

**RQ 1.3**

What are the challenges medical models face at document-level?

### 3.4.1 Case Study VI: How much to read to understand?

This chapter is based on work previously published in our article: *Sinha, A., Bigeard, S., Clausel, M., & Constant, M. (2023). What shall we read: the article or the citations?-A case study on scientific language understanding. In Actes de CORIA-TALN 2023. Actes de l'atelier "Analyse et Recherche de Textes Scientifiques" (ARTS)@TALN 2023, pages 80–85, Paris, France. ATALA.* Parts of the text, figures, and results are adapted from this publication.

Scientific databases[21] are accumulating a large amount of literature to such an extent that it is getting overwhelming (Johnson et al., 2018) and practically impossible to remain up-to-date for researchers (See fig. 3.18). Several recent works have looked into building intelligent systems for tasks such as ad-hoc based retrieval, conversational agents, recommendation, summarization, document search, and re-ranking, as shown by the advances in the area of Neural Information Retrieval (Neural IR) and Biomedical Text Mining (Zhang et al., 2016; Gu et al., 2021b; Thakur et al., 2021).

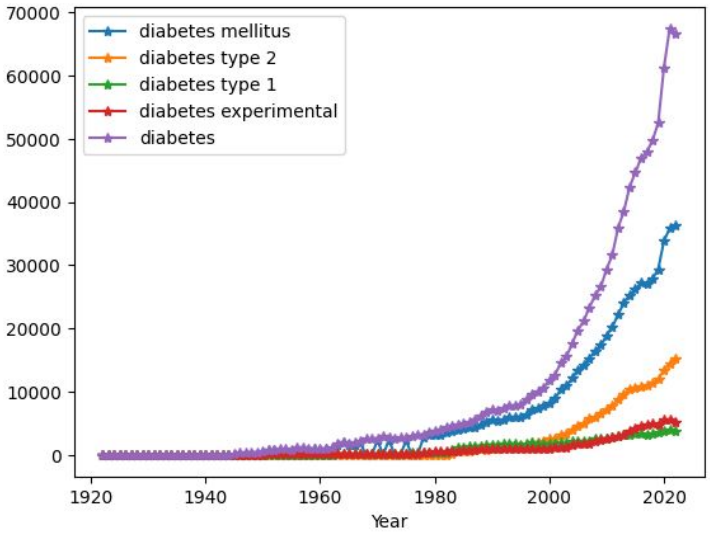



Figure 3.18: Papers containing related to Diabetes according to PubMED until June 2023.

Given the unstructured nature of scientific documents, their lengthiness and the presence of metadata, the key question that arises when dealing with scientific information extraction is how much data can be helpful to ..... ?

**Research Question 1.2c:** *How much text do LMs require to capture the relevant information contained in [SCI] medical data?*

We investigate the informativeness of different types of *text-based* features in a the

[21]Examples: Pubmed (https://pubmed.ncbi.nlm.nih.gov/), ASCO (https://www.asco.org/); ArXiv (https://arxiv.org/), etc.

document classification setting. Additionally, we further explore the impact of increasing the amount of text features. This work investigates the impact of different text features for scientific document classification task. The effectiveness of more text for text-ranking and text-retrieval tasks has been studied previously (Lin, 2009). Several works (Guo et al., 2011; Ermakova et al., 2018; Yeganova et al., 2021) have also investigated the importance of features from different sections in scientific articles using methods such as traditional BM25 scoring, deep-NN (Huang et al., 2013) and weighted word-count (Yu et al., 2014). In parallel, transfer learning (Peng et al., 2019; Gu et al., 2021b; Kanakarajan et al., 2021) has received a lot of attention with domain specific pre-trained language models (PLMs) for various downstream tasks such NER, relation extraction, question answering, document classification for scientific and biomedical text mining.

## Methodology

**Dataset** We use the *PubMed-Diabetes* (Namata et al., 2012) for our experiments. The dataset contains 19717 articles belonging to 3 classes of diabetes-mellitus, *Experimental*, *Type-1*, and *Type-2*.

**Models.** We consider BERT (Devlin et al., 2019) Classification Token [CLS] (CLS) as our baseline. Then, we implemented a Text-CNN model (Zhang and Wallace, 2015), more concretely, a BERT-CNN architecture. This model applies five stacked 2D-CNN blocks to the token-level representations from BERT, followed by a max-pooling layer, in order to capture richer local and hierarchical textual features.

As a control we include a random noise feature to highlight the effect of incorporating meaningful text into the models.

For lexical features, we use Term Frequency-Inverse Document Frequency (TF-IDF) representations in two forms: (i) 500-dimensional vectors provided with the dataset, based on a curated keyword list (Namata et al., 2012) (referred to as `tfidf-c`); and (ii) manually generated TF-IDF vectors (`tfidf-m`) computed from specific text portions (e.g., title-only or abstract-only) for comparative analysis.

To capture contextual information, we employ BERT (Devlin et al., 2019), using the [CLS] token embedding from the model pretrained on English Wikipedia and BookCorpus as a general-domain baseline. For domain specific models, we consider several pretrained language models (PLMs): BioBERT (Lee et al., 2020), which is further pretrained on large-scale biomedical corpora (PubMed and PMC[22]); and PubMedBERT (Gu et al., 2021b), which, unlike BioBERT, is trained from scratch solely on PubMed abstracts. We also include BioELECTRA (Kanakarajan et al., 2021), a biomedical adaptation of ELECTRA (Clark et al., 2020), pretrained from scratch on PubMed and PMC using the PubMedBERT vocabulary. BioELECTRA has shown superior performance over prior biomedical PLMs, achieving state-of-the-art results on the BLURB

[22]PMC stands for PubMed Central.

benchmark (Gu et al., 2021b).

**Implementation Details** In our experiments, we follow Yang et al. (2016)s' the data splitting method, keeping 20 random instances of each class for training, 500 random instances for development set, and 1000 constant instances for test. We report the average results for 5 seeded model runs. We keep the hyperparameter setting common across all the models. We use a learning-rate of 1e-06, train for 500 epochs 500 with a max length for title=64 and abstract=512, and batch-size of 64 for title and 8 for abstract.

## Results & Discussion

Table 3.21 shows the results divided into two subsections. The first subsection shows *random* features to compare and show the impact of feature-based models. In the next subsection, we have *text-only* based features including TF-IDF and PLM-based models with two variants: [CLS]-finetuning and Text-CNN model.

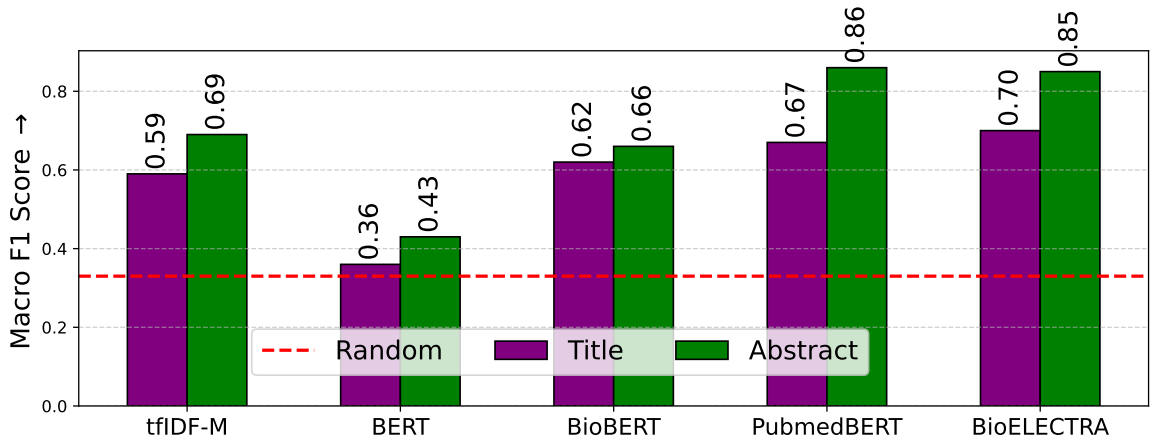



Figure 3.19: Performance comparison of language (embedding) models for Scientific Document Classification.

We notice that *tfidf-c* performs better than *tfidf-m*, which can be attributed to the curated keyword list compared as automatic generated keyword list. We continue with *tfidf-m* to study **Title**- and **Abst**- text separately. We observed that PLM-based features perform predominately better than *tfidf* features with the exception of `BERT-base-case` (because of its general vocabulary) and further, all the **Abst**- models outperform the **Title**- based models showing the relevance of the Abstract section. We notice that surprisingly BioElectra performs lower than PubMedBERT for all cases (except for [CLS] baseline with title corpora) and its performance degrades with Text-CNN model.

## Summary

In this case study, we present a benchmarking analysis to evaluate how the amount of available textual input influences the performance of various language (embedding) models in scientific document classification tasks. Our experiments compare traditional feature-based methods such as term frequency-inverse document frequency (TF-IDF) with several transformer-based models, including BERT, BioBERT, PubMedBERT, BioELECTRA, and a CNN-augmented variant of BERT for document encoding.

The empirical results validate the intuitive hypothesis that model performance generally improves with longer input text. This trend is particularly evident for transformer-based models, which are known to benefit from richer contextual information. The study also highlights the potential value of incorporating additional modalities beyond textual features such as citation graphs and structured metadata especially within the scientific domain (SCI), where such information is often readily available but underutilized.

These findings underscore the importance of considering both the quantity and structure of input information when designing NLP systems for scientific document understanding. They further suggest promising directions for enhancing classification performance through multimodal or hybrid approaches that integrate textual and non-textual signals.

**Limitations** A key limitation of this study lies in the use of only abstracts rather than full-text articles. While abstracts provide a concise summary, they may omit important contextual and methodological details found in the body of the article. Including full texts in future experiments could reveal additional insights, especially considering the frequent presence of redundancy, complex discourse structures, and long-range dependencies that are better captured with extended context.

| | Method | Title | | | | Abstract | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **P** | **R** | **F1** | **Acc** | **P** | **R** | **F1** | **Acc** |
| R | random | $0.36_{0.03}$ | $0.33_{0.03}$ | $0.33_{0.03}$ | $0.32_{0.03}$ | - | - | - | - |
| T | tfidf-c | $0.70_{0.02}$ | $0.68_{0.03}$ | $0.67_{0.03}$ | $0.68_{0.03}$ | - | - | - | - |
| | tfidf-m | $0.61_{0.03}$ | $0.60_{0.02}$ | $0.59_{0.03}$ | $0.60_{0.02}$ | $0.71_{0.02}$ | $0.70_{0.03}$ | $0.69_{0.03}$ | $0.70_{0.03}$ |
| | BERT | $0.41_{0.09}$ | $0.43_{0.02}$ | $0.36_{0.04}$ | $0.43_{0.02}$ | $0.45_{0.16}$ | $0.50_{0.07}$ | $0.43_{0.11}$ | $0.50_{0.07}$ |
| | BioBERT | $0.67_{0.04}$ | $0.64_{0.06}$ | $0.62_{0.09}$ | $0.64_{0.06}$ | $0.72_{0.08}$ | $0.67_{0.14}$ | $0.66_{0.17}$ | $0.67_{0.14}$ |
| | PubmedBERT | $0.71_{0.06}$ | $0.68_{0.07}$ | $0.67_{0.09}$ | $0.68_{0.07}$ | $0.86_{0.02}$ | $0.86_{0.02}$ | $0.86_{0.02}$ | $0.85_{0.03}$ |
| | BioElectra | $0.76_{0.03}$ | $0.71_{0.04}$ | $0.70_{0.05}$ | $0.71_{0.04}$ | $0.86_{0.02}$ | $0.85_{0.02}$ | $0.85_{0.02}$ | $0.84_{0.03}$ |
| | BERTCNN | $0.60_{0.07}$ | $0.59_{0.08}$ | $0.59_{0.08}$ | $0.59_{0.08}$ | $0.66_{0.05}$ | $0.65_{0.05}$ | $0.65_{0.05}$ | $0.65_{0.05}$ |
| | BioBERTCNN | $0.70_{0.06}$ | $0.69_{0.06}$ | $0.69_{0.06}$ | $0.69_{0.06}$ | $0.72_{0.03}$ | $0.70_{0.03}$ | $0.70_{0.04}$ | $0.69_{0.05}$ |
| | PubmedCNN | $0.83_{0.03}$ | $0.82_{0.04}$ | $0.82_{0.04}$ | $0.82_{0.04}$ | $0.87_{0.03}$ | $0.87_{0.03}$ | $0.87_{0.04}$ | $0.87_{0.04}$ |
| | BioElectraCNN | $0.65_{0.10}$ | $0.62_{0.06}$ | $0.58_{0.09}$ | $0.62_{0.06}$ | $0.85_{0.02}$ | $0.84_{0.02}$ | $0.84_{0.02}$ | $0.84_{0.03}$ |

Table 3.21: Complete Experiment results.

### 3.4.2 Case Study VII: Can LMs learn with limited data?

This chapter is based on work previously published in our article: *Sinha, A., & Buhnila, I. (2025, June). ATILF at NTCIR-18 RadNLP 2024 Shared Task: With less radiology reports, comes less performance. In NTCIR-18: Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies (pp. 354-358).* Parts of the text, figures, and results are adapted from this publication.

A major challenge in medical NLP is the limited availability of high-quality annotated datasets, which makes it difficult to directly translate the performance of language models on standard language understanding benchmarks to clinical settings. Existing strategies to address this low-resource constraint include augmenting the available datasets using text augmentation techniques (Collado-Montañez et al., 2025) or leveraging additional publicly available corpora for pre-training (Zhang et al., 2024a).

However, even with these adaptations, language models, and in particular large language models, are prone to generating factually incorrect information a phenomenon known as hallucination (Huang et al., 2025) which can be particularly dangerous in mission-critical setting such as medicine. Recent approaches, such as Retrieval-Augmented Generation (RAG) (Gao et al., 2023), have shown promise in mitigating hallucinations by integrating retrieval mechanisms and self-reflection, thereby improving factual accuracy and reliability in medical reasoning tasks (Lewis et al., 2020b; Asai et al., 2023; Jeong et al., 2024).

In this study, we consider two clinical classification tasks[23] : (i) Automated Lung Cancer Staging; and (ii) Detection of insomnia in clinical notes.

In both the datasets, we are presented with a common challenge i.e., too little data to learn a medical phenomena from. Using these dataset, we are interested to our investigate the following research question:

**Research Question 1.2d:** *To what extent LMs can learn from small [CLIN] medical datasets?*

## Methodology

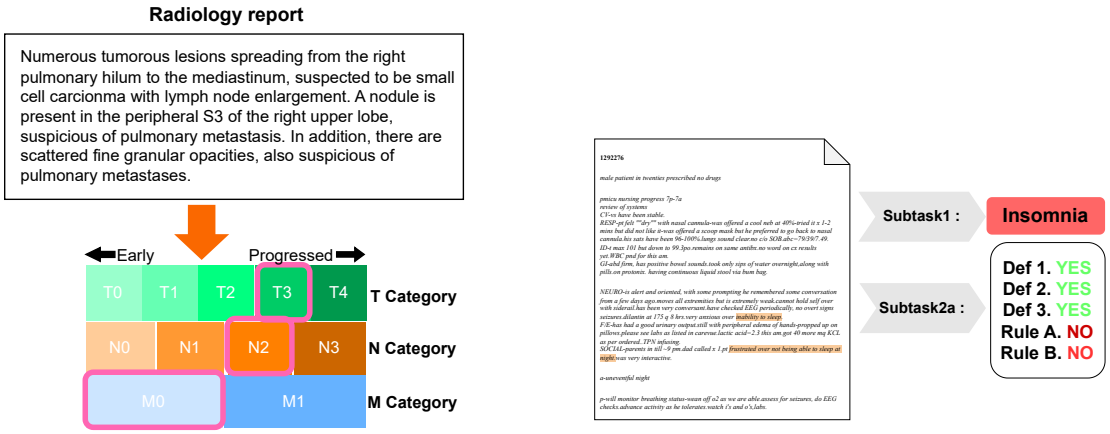
**Dataset.** The Automatic clinical Staging dataset is part of RadNLP challenge (Nakamura et al., 2024). It contains two languages : en and jp, where the en version of the dataset in a machine translated version of jp dataset. And, for the detection of insomnia in clinical notes, we use a col-

| | TRAIN | DEV | TEST |
|---|---|---|---|
| RADNLP | 108 | 54 | 81 |
| INSOMNIA | 70 | 20 | 100 |

Table 3.22: Dataset Statistics.

[23] *Automated Lung Cancer Staging* was introduced as part of the NTCIR-17 Medical Natural Language Processing for Social Media and Clinical Texts (MedNLP-SC) shared task (Nakamura et al., 2023). And, *detection of insomnia in clinical notes* was introduced as part of the SMM4H-HeaRD 2025 shared tasks[24].

lection of electronic health records (EHRs) from MIMIC-III Clinical Database (v1.4) Johnson et al. (2016) that was provided by the organizer of SMM4H Task 4 shared task[25] . The clinical reports are annotated by experts of symptoms of Insomnia. The dataset statistics is shown in Table 3.22.




(a) Automatic Clinical Staging Classification.

(b) Insomnia Classification for Clinical Notes.

Figure 3.20: Illustration of the two classification tasks.

**Task Description.** (a) The *Automatic Clinical Staging*[26] task addresses the automatic determination of the clinical stage[27] of lung cancer directly from radiology reports. Concretely, the problem is framed as a multi-class classification task spanning the three tumor-related dimensions defined in the Tumor, Node, Metastasis Cancer Staging (TNM) system, where each of these dimensions is further subdivided into fine-grained classes, making the classification problem highly detailed and clinically meaningful. The TNM classification system is used to stage cancer based on three components: **T (Tumor)** refers to the size and/or local extension of the primary lesion, with ten possible subclasses (T0, Tis, T1mi, T1a, T1b, T1c, T2a, T3, T4); **N (Node)** denotes the presence and extent of regional lymph node metastasis and includes four subclasses (N0, N1, N2, N3); and **M (Metastasis)** assesses the presence and extent of distant metastasis, categorized into four subclasses (M0, M1a, M1b, M1c).

(b) The *Insomnia Detection task* contains two subtasks: The first subtask involves a binary classification for identifying the clinical note for the patient suffering with the condition of Insomnia. The second subtask is a multilabel classification task which involves evaluating each clinical note against clinically defined Insomnia rules (see table 3.24): Definition 1, Definition 2, Rule A, Rule B, and Rule C.

[25]https://github.com/guilopgar/SMM4H-HeaRD-2025-Task-4-Insomnia

[26]Clinical staging reflects the degree of disease progression and plays a pivotal role in guiding treatment decisions and estimating patient prognosis.

[27]The staging criteria follow the 8th edition of the TNM Classification of Malignant Tumors, as defined by the Union for International Cancer Control (UICC). Table 3.23 provided the definitions from https://radiologyassistant.nl/chest/lung-cancer/tnm-classification-8th-edition-1

| Type | Definition |
|---|---|
| T0 | No primary tumor |
| Tis | Ground-glass nodule without solid component with the total diameter $\leq$3 cm |
| T1mi | Ground-glass nodule with solid component $\leq$0.5 cm and the total diameter $\leq$3 cm |
| T1a | Solid component diameter $\leq$1 cm |
| T1b | Solid component diameter $>$1 cm and $\leq$2 cm |
| T1c | Solid component diameter $>$2 cm and $\leq$3 cm |
| T2a | Solid component diameter $>$3 cm and $\leq$4 cm. Otherwise, extension to main bronchus or visceral pleura, or atelectasis or obstructive pneumonia extending to hilum, with the solid component diameter $<$3 cm or unknown |
| T2b | Solid component diameter $>$4 cm and $\leq$5 cm |
| T3 | Solid component diameter $>$5 cm and $\leq$7 cm. Otherwise, solid component diameter $\leq$5 cm and either condition holds: direct invasion of parietal pleura, chest wall (including superior sulcus tumor), mediastinal nerve, or pericardium; separate tumor nodule(s) in the same lobe |
| T4 | Solid component diameter $>$7 cm. Otherwise, either condition holds: invasion of diaphragm, mediastinum, heart, great vessels, trachea, recurrent laryngeal nerve, esophagus, spine, or carina; tumor nodule(s) in a different ipsilateral lobe |
| N0 | No regional lymph node metastasis |
| N1 | Metastasis to ipsilateral peribronchial, hilar, or pulmonary lymph nodes, including direct invasion of the primary tumor |
| N2 | Metastasis to ipsilateral mediastinal or subcarinal lymph nodes |
| N3 | Metastasis to contralateral mediastinal, hilar, anterior scalene, or supraclavicular lymph nodes |
| M0 | No distant metastasis |
| M1a | Contralateral tumor nodule(s), pleural or pericardial nodule(s), malignant pleural effusion, or malignant pericardial effusion |
| M1b | Single extrathoracic metastasis |
| M1c | Multiple extrathoracic metastases |

Table 3.23: TNM classification mapping with KB_radnlp definitions from tnm-classification-8th-edition.

| Definition 1 | Definition 2 | |
|---|---|---|
| *Trouble initiating sleep* | *Fatigue or malaise* | *Behavioral problems such as hyperactivity, impulsivity, or aggression* |
| *Trouble maintaining sleep* | *Impaired attention, concentration, or memory* | *Decreased motivation, energy, or initiative* |
| *Waking up earlier than desired* | *Impaired social, family, occupational, or academic performance* | *Proneness to errors or accidents* |
| *An explicit mention of Insomnia* | *Mood disturbance or irritability* | *Concerns or dissatisfaction with sleep* |
| | *Daytime sleepiness* | |
| **Rule A** | *The patient has insomnia if they meet both **Definition 1** and **Definition 2*** | |
| **Rule B** | *If the patient is prescribed any of the following meds: Estazolam, Eszopiclone, Flurazepam, Lemborexant, Quazepam, Ramelteon, Suvorexant, Temazepam, Triazolam, Zaleplon, Zolpidem* | |
| **Rule C** | *If the patient is prescribed any of the following meds: Acamprosate, Alprazolam, Clonazepam, Clonidine, Diazepam, Diphenhydramine, Doxepin, Gabapentin, Hydroxyzine, Lorazepam, Melatonin, Mirtazapine, Olanzapine, Quetiapine, Trazodone. OR any symptoms from **Definition 1** or **Definition 2*** | |

Table 3.24: Description of the Insomnia Detection in Clinical Notes.

**Models.** For Automatic Clinical Staging, we consider two pool of models: Small Language Model (SLM) — a general language SLM Llama3.2-3B (Dubey et al., 2024), and a pool of pre-trained language models (PLM), BioBERT (Lee et al., 2020) and Bio-ClinicalBERT (Alsentzer et al., 2019b).

For Insomnia Detection, we considered a pool of pre-trained language models including BioBERT (Lee et al., 2020), PubmedBERT (Gu et al., 2021a), SciBERT (Beltagy et al., 2019), MedBERT (Vasantharajan et al., 2022), ClinicalBigbird (Li et al., 2022) for the first subtask. For second subtask, for Definition 1 and 2, we used the above mentioned language models, however, for Rule A we used logical operator module and finally, we used a fuzzy string matching module for Rule B and C.

[INST]
Given this radiology report : [`rr_text`],

classify it to [`var`]-clinical staging based on [`desc`], provided your options are as follows: `option_list`. **GIVE ONLY THE CORRECT ANSWER** (Pick only one option).
/[/INST]

Figure 3.21: Default prompt for Automatic Clinical Staging Task

**Experimental Setup.** For Automatic Clinical Staging, we detail our approach across three configurations:

1. First, we use a prompt-based LLM baseline, where the input consists of a task-specific prompt concatenated with the radiology report (see Figure 3.21).
2. Second, we enhance the prompt by including TNM classification tags (e.g., *'T0', 'Tis', 'T1mi'*) along with their definitions, such as $t_{\text{desc}}$ = "Assessment of the size and/or extension of the primary lesion" (see Figure 3.25), using the knowledge base described in Table 3.23. Both the vanilla and definition enhanced prompt based prediction were post-processed[28].
3. Finally, we train PLM classifiers under a standard configuration for 20 epochs, considering two setups: one with separate training for T, N, and M classifications, and another with joint training over all possible TNM class permutations.

For Insomnia Detection, we performed supervised training for all the models with focal loss (Lin et al., 2017) to address the class imbalance. Focal Loss enables the standard cross-entropy loss designed to address class imbalance by focusing on "hard" examples, while reducing the loss for "easy" examples.

$$L_{\texttt{focal}} = -\alpha_t(1-\hat{p}_{i,y_i})^{\gamma}\texttt{log}(\hat{p}_{i,y_i})$$

where $\gamma$ is the focusing parameter to reduce the contribution of "easy" examples (default set to 2.0), and $\hat{p}_{i,y_i}$ is the softmax probability for the correct class $y_i$.

**Evaluation Metrics.** For automatic clinical staging task, Accuracy was used as the metrics with Fine and Coarse level of evaluation. In Fine Accuracy, exact class is matched and Coarse Accuracy ignores the distinctions between Tis/T1mi/T1a/T1b/T1c, T2a/T2b, and M1a/M1b/M1c. And for Insomnia Detection task, for the first subtask, micro-average F1 Score was used but the subtask 2A, an average over all the rules evaluation.

# Results & Discussion

**Automatic Clinical Staging.** Figure 3.22 presents the results on validation set (left) and test set (right). Table 3.26 presents the complete validation and test scores for all the models. Overall, clinical staging task remained challenging for all language models. The results suggested that

| | Classes | Effec. CIR (Yu et al., 2022) |
|---|---|---|
| T | 10 | 10.33 |
| N | 4 | 5 |
| M | 4 | 5.28 |
| TNM | 160 | 12 |

Table 3.25: Class Imbalance Ratio of Automatic Clinical Staging Datasets.

[28]For both vanilla and definition-enhanced prompting, we first performed automatic extraction of predictions from the raw LLM outputs, followed by a minimal manual review to discard predictions that were entirely irrelevant or unusable.

out of the pool of models on the validation set
BioClinicalBERT obtained the hightest score with a fine-grained joint accuracy of 14.8 and coarse joint accuracy of 27.78. On test set, BioClinicalBERT obtained 28.4 joint accuracy (Fine) however there was significant gap of 37% in comparison to the best submission received for the challenge (Nakamura et al., 2024) which was based on model tuning, ensembling, and data augmentation method.

Based on the results, for individual category of clinical staging, tumor size/extension classification (T) class was the most difficult task for all the language models which is due to its highest CIR score based on the train set. The best results on the T class fine and coarse accuracy were obtained with Llama3.2+Definition on fine grained accuracy (0.2592) followed by BioClinicalBERT on coarse accuracy (0.3889). For the extent of lymph node metastasis class (N class), BioBERT obtained the highest accuracy score of 0.44 which was also the highest score among the pools of pre-trained language models considered in our study for automatic clinical staging. This finding is aligned with intuition as the the effective CIR score the lowest for N. The three clinical PLM models were the most efficient on this class, while Llama3.2, a general model, achieved a lower accuracy scores (0.25). And finally, the M class (extent of distant metastasis) was the easiest for all language models tested in terms of coarse accuracy due to its binary categorization for coarse setting.

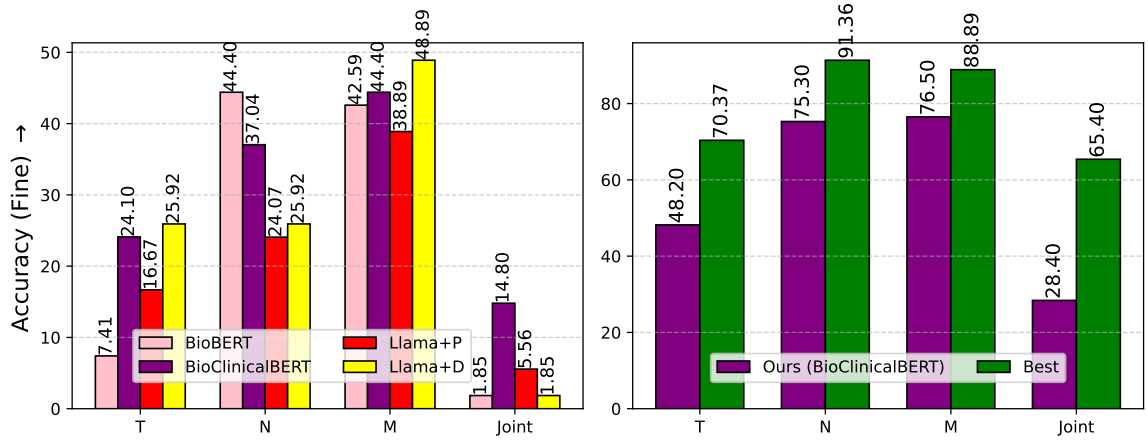



Figure 3.22: Validation (left) and Test (right) Results comparison for Automatic Clinical Staging Task.

Additionally, we observed an improvement in the fine-grained scores for the individual clinical staging classes when definitions were provided in the prompt. However, this didn't help in the case of Joint Accuracy due to lack of clear definition for each of 160 joint classes together. Complementary lexical and semantic information with the definitions helped language models perform better semantic disambiguation to operationalize the individual T, N, M classification task.

**Insomnia Detection.** Figure 3.23 presents results on validation set and fig. 3.24 provides an overall comparison of all the systems submitted at the shared task aggregated

as PLMs and LLMs. Firstly, for direct insomnia classification task (Subtask 1), we notice Clinical Bigbird, SciBERT and MedBERT outperform other models, this can be attributed to their pretraining datasets which in case of clinical BigBIRD is MIMIC-III and for the latter is n2c2, BioNLP, and CRAFT community datasets. For the rule-based classification (subtask 2A), we notice an overall lower performance compared to direct classification, this can be attributed to the complexity of the task as it involves in investigating 5 rules together (See table 3.24). Other factor, that can be attributed is less number of training samples, in specifically for Definition 2 . Further, as Rule A and C are dependent of Definition 1 and 2, they are directly effected by models' performance on each of the respectively.

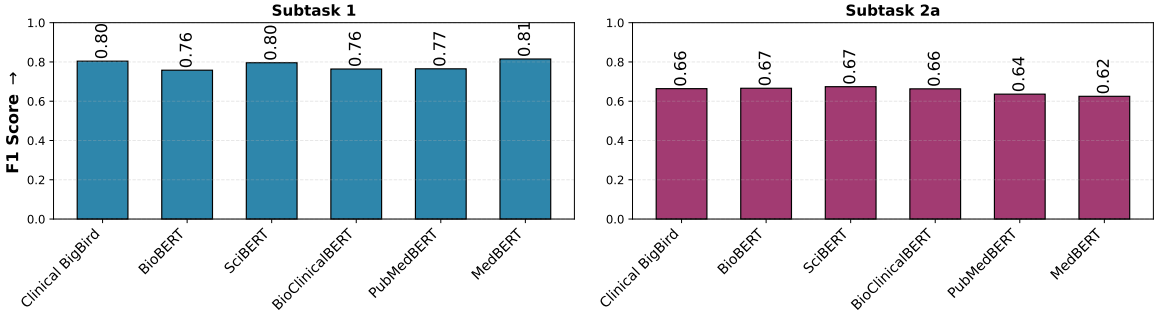



Figure 3.23: Insomnia Resuls on Validation Set.

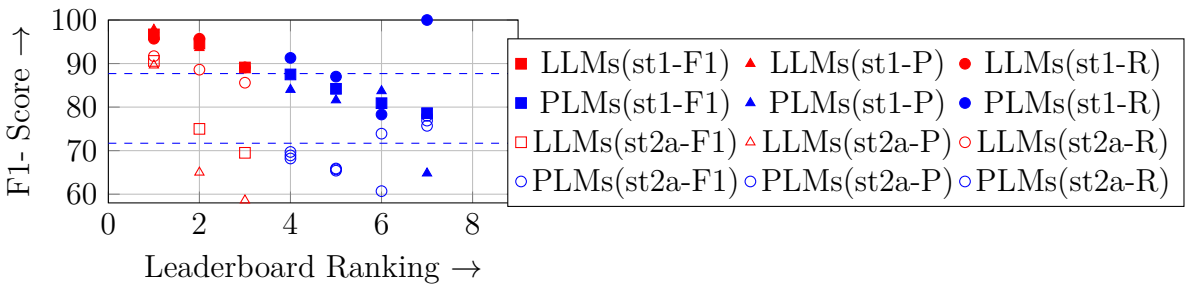



Figure 3.24: Comparison of PLMs versus LLMs for Insomnia Detection.

During error analysis,we identified two samples in the validation set for which all the PLM models (See table 3.27, Subtask 1) consistently failed to predict the correct class for which we evaluated GPT-4o and obtained 100% agreement with the gold labels. We found out that phrases such as "*very tired and weak*", "*nightmares*" in the first example and "*neuro / short deficit*", "*very cooperative, sweet, now at times angry*" and "*discomfort and incisional pain*" in second sample (see table 3.29) can loosely be associated with Insomnia. These findings suggests that the models often marked even any non-specific symptoms even if it's not strictly constrained to explicit mentions or direct evidence as an indication of Insomnia. As shown in table 3.27 and 3.28, the `recall(R)` of all the PLMs is $\geq$ 85%, implying the presence of false-positives. This

indicates that the PLMs are more cautious and tend to prioritize identifying most or all actual cases of insomnia, even at the cost of precision.

For subtask 2A, the subtask is composed of multiple sub-components (Definition 1, Definition 2, Rule A, Rule B and Rule C) which have a non-zero overlap (See table 3.24) but are evaluated collectively. As a result, an incorrect prediction in any of the sub-components can lead to an overall drop in performance. This is evident in the lower overall performance compared to Subtask 1, which can also be attributed to the complexity of Subtask 2A as it involves identifying symptoms in the clinical notes which not necessarily are mentioned explicitly. Furthermore, since Rule A and C are dependent on Definition 1 and 2, they are directly effected by models' performance on each of the respectively.

Overall, for both the classification, automatic clinical staging and insomnia detection classification we observe while the pre-trained language models under the minimal setting did not surpass sophisticated large language model setup by other participating teams. Considering the scarcity of data setting, PLM based modeling provides more for improvement importantly considering the computational intensive nature of LLM based approaches.

## Summary

In this study, we investigated the effectiveness of pretrained language models (PLMs) and large language models (LLMs) for two clinical text classification tasks: TNM clinical staging and insomnia detection. Our findings suggest that, in low-resource settings, domain-specific PLMs are generally more effective, demonstrating competitive performance without requiring large amounts of training data. However, their effectiveness remains limited in the absence of additional knowledge sources such as structured knowledge bases or extensive annotated datasets.

While domain-specific PLMs show promise, their performance does not consistently scale to match that of general-purpose LLMs, particularly in specialized clinical contexts. This highlights the limitation of relying solely on pretraining in the domain: the generalization capabilities of PLMs are insufficient without task-specific or external context. Conversely, LLMs demonstrate stronger performance in scenarios where supplementary information such as class definitions or structured prompts is provided, as seen in the clinical staging classification task.

**Limitations** A key limitation of this study is the lack of systematic exploration of advanced techniques for leveraging LLMs, such as instruction tuning, few-shot prompting, or in-context learning with tailored prompts. Future work should investigate the impact of such techniques to better assess the full potential of LLMs in clinical NLP tasks.

| Model | Joint Acc. (Fine) | T Acc. (Fine) | N Acc. (Fine) | M Acc. (Fine) | Joint Acc. (Coarse) | T Acc. (Coarse) | N Acc. (Coarse) | M Acc. (Coarse) |
|---|---|---|---|---|---|---|---|---|
| Llama3.2+P | 0.0556 | 0.1667 | 0.2407 | 0.3889 | 0.0926 | 0.3519 | 0.2407 | 0.5740 |
| Llama3.2+D | 0.0185 | 0.2592 | 0.2592 | 0.4814 | 0.0555 | 0.3333 | 0.2592 | 0.5556 |
| BioClinicalBERT | 0.1481 | 0.2407 | 0.3704 | 0.4444 | 0.2778 | 0.3889 | 0.3704 | 0.5000 |
| BioBERT | 0.0185 | 0.0741 | 0.4444 | 0.4259 | 0.2037 | 0.2778 | 0.4444 | 0.4815 |

(a) Maintask; (P: Prompting, D: Definitions)

| Model | Joint Acc. (Fine) | T Acc. (Fine) | N Acc. (Fine) | M Acc. (Fine) | Joint Acc. (Coarse) | T Acc. (Coarse) | N Acc. (Coarse) | M Acc. (Coarse) |
|---|---|---|---|---|---|---|---|---|
| Best | 0.6543 | 0.7037 | 0.9136 | 0.8889 | 0.6914 | 0.7407 | 0.9136 | 0.9136 |
| Mean | 0.4105 | 0.5277 | 0.7955 | 0.7646 | 0.4976 | 0.5972 | 0.7955 | 0.8202 |
| Median | 0.5247 | 0.6172 | 0.9012 | 0.8395 | 0.5617 | 0.6605 | 0.9012 | 0.8827 |
| Ours | 0.284 | 0.4815 | 0.7531 | 0.7654 | 0.4815 | 0.6296 | 0.7531 | 0.9012 |

(a) Maintask

Table 3.26: Detailed Performance Comparison on Automatic Clinical Staging Task.

| | **Subtask 1** | | | **Subtask 2a** | | |
|---|---|---|---|---|---|---|
| model | F1 | P | R | F1 | P | R |
| Clinical BigBird | 0.804 | 0.704 | 0.950 | 0.664 | 0.824 | 0.557 |
| BioBERT | 0.758 | 0.680 | 0.867 | 0.666 | 0.771 | 0.590 |
| SciBERT | 0.796 | 0.706 | 0.917 | 0.674 | 0.722 | 0.648 |
| BioClinicalBERT | 0.764 | 0.687 | 0.867 | 0.663 | 0.747 | 0.609 |
| PubMedBERT | 0.765 | 0.699 | 0.850 | 0.636 | 0.716 | 0.581 |
| MedBERT | 0.815 | 0.728 | 0.933 | 0.625 | 0.639 | 0.623 |

Table 3.27: Results on validation set for Insomnia Detection Task for direction classification (on left; Subtask 1) and rule-based classification (on right; Subtask 2A) .

| | **Subtask 1** | | | **Subtask 2a** | | |
|---|---|---|---|---|---|---|
| | F1 | P | R | F1 | P | R |
| MEAN | 0.877 | 0.853 | 0.913 | 0.717 | 0.673 | 0.788 |
| MEDIAN | 0.869 | 0.840 | 0.935 | 0.692 | 0.650 | 0.818 |
| MIN | 0.786 | 0.648 | 0.761 | 0.607 | 0.521 | 0.515 |
| MAX | 0.967 | 0.978 | 1.000 | 0.906 | 0.896 | 0.932 |
| | *Our submissions* | | | | | |
| Ens. (Top 3) | 0.868 | 0.811 | **0.935** | 0.681 | 0.646 | 0.719 |
| Ens. (Top 5) | 0.868 | 0.811 | **0.935** | **0.689** | **0.681** | 0.697 |
| Ens. (Top 10) | **0.875** | **0.840** | 0.913 | 0.681 | 0.640 | **0.727** |

Table 3.28: Detailed results for Insomnia Detection Classification task.

| | **Clinical Report** | Insomnia | GPT |
|---|---|---|---|
| | Subtask 1 | | |
| 10 | female patient in eighties prescribed Digoxin, Hydromorphone, Propranolol LA, Sodium Chloride 0.9% Flush, Heparin, Levofloxacin, Vancomycin HCl, Iso-Osmotic Dextrose, Dextrose 5%, Furosemide, NS, Metronidazole, Potassium Chloride, Acetaminophen, Metoprolol, Magnesium Sulfate, Miconazole Powder 2% mental status: alert orinted. obeys commands.pt slept most of night. cv: bp up to 175-180 with discomfort.bp decrease to 120-130 when conmfortable.,hr perm pacer av pacing at 72. gu: urine output low. 20-25 cc/hr. urine ouput dropped off to 10 cc about 0200. treated with 500 cc ns urine picked up to 20-25 cc/hr gi: pos bowel sounds abd soft. small amount of soft stool. integumentary: buttocks and r breast very pink rash. needs mycostatin .spoke with team and asked for order for mycostatin powder. right breast is macerated..skin is peeling. resp: pt has clear upper airways ,diminished at bases. coughing and raising small amounts white.o2 at 2 liters nc, chest tube draining sm amounts serosanguinous. pt is very tired and weak. she is uncomfortable but does no want any med stronger thatn tylenol.because she says it gives her nightmares. tylenol times 2 with fair effect. | no | no |
| 12 | female patient in sixties prescribed no drugs 7a-7p NPN s: I hurt everywhere o: see carevue for all objective data neuro: pt c/o no visual disturbances, MAE, strength =, pt w/ some short term deficits. Family states pt at baseline is very cooperative, sweet, now at times angry, but is cooperative. Neuro here for consult. Head CT done. cv: hemodynamically stable w/ hr 70-80's sr, pacing wires attached. bp 130-160/50-70. resp: sats 92-95 on RA, lungs w/ crackles at bases. id: tm 99.0 po heme: hct 23.4 this am, HO aware. end: bs 130-234, covered per riss. gu: foley draining cl yel urine 100cc/hr gi: taking sm amts soft food, no stool today. skin: chest and leg incisions d/i pain: c/o general discomfort and incisional pain, given tylenol and percocet w/ relief. activity: oob to chair w/ 2 assista, walked 30' w/ PT using wheeling walker, around rm w/ rn. social: 3 daughters in to visit, met w/ social worker [**Name (NI) **] [**Last Name (NamePattern1) **], case manager [**Name (NI) 346**] [**Last Name (NamePattern1) **], Dr [**Last Name (STitle) 10415**] in to speak w/ family and pt. A: neuro changes s/p CABG P: Monitor neuro status, goal keep sbp >130. Continue cardiac rehab. | no | no |

Table 3.29: Examples from Task 4 (Subtask 1) validation set where every PLM failed to predict the correct class. Highlighted words denote the reasoning for PLM's decision for mislabelling the clinical report for Insomnia by GPT.

The TNM staging system is a method for classifying the extent of cancer spread. It stands for Tumor, Node, and Metastasis, and is used to describe the size of the primary tumor (T), the extent of lymph node involvement (N), and the presence of distant metastasis (M).

Tumor (T): Describes the size and extent of the primary tumor. The classes are:

['T0|No primary tumor', 'Tis|Ground-glass nodule without solid component with the total diameter ≤3 cm', 'T1mi|Ground-glass nodule with solid component ≤0.5 cm and the total diameter ≤3 cm', 'T1a|Solid component diameter ≤1 cm', 'T1b|Solid component diameter >1 cm and ≤2 cm', 'T1c|Solid component diameter >2 cm and ≤3 cm', 'T2a|Solid component diameter >3 cm and ≤4 cm. Otherwise, extension to main bronchus or visceral pleura, or atelectasis or obstructive pneumonia extending to hilum," with the solid component diameter <3 cm or unknown', 'T2b|Solid component diameter >4 cm and ≤5 cm', 'T3|Solid component diameter >5 cm and ≤7 cm. Otherwise, solid component diameter ≤5 cm and either condition holds: direct invasion of parietal pleura, chest wall (including superior sulcus tumor), mediastinal nerve, or pericardium; separate tumor nodule(s) in the same lobe', 'T4|Solid component diameter >7 cm. Otherwise, either condition holds: invasion of diaphragm, mediastinum, heart, great vessels, trachea, recurrent laryngeal nerve, esophagus, spine, or carina; tumor nodule(s) in a different ipsilateral lobe']

Node (N): Describes the degree of spread to regional lymph nodes. The classes are:

['N0|No regional lymph node metastasis', 'N1|Metastasis to ipsilateral peribronchial, hilar, or pulmonary lymph nodes, including direct invasion of the primary tumor', 'N2|Metastasis to ipsilateral mediastinal or subcarinal lymph nodes', 'N3|Metastasis to contralateral mediastinal, hilar, anterior scalene, or supraclavicular lymph nodes']

Metastasis (M): Indicates whether the cancer has spread to distant parts of the body. The classes are:

['M0|No distant metastasis', 'M1a|Contralateral tumor nodule(s), pleural or pericardial nodule(s), malignant pleural effusion, or malignant pericardial effusion', 'M1b|Single extrathoracic metastasis', 'M1c|Multiple extrathoracic metastases']

I want you to read the radiology report and assess it for clinical staging. Give me an answer for T, one for N, and one for M only using the classes mentioned above. Give only the answer, no additional text. Use this json format to answer 'T':" , 'N':" , 'M':" ; radiology report = ###

Figure 3.25: Prompt for Automatic Clinical Staging enhanced with definitions

## 3.5 Temporal problems

Understanding temporal information is essential in clinical setting. Events such as symptom onset, diagnoses, treatments, and outcomes are inherently time-dependent, and their interpretation relies on knowing when they occurred and in what order (Zhou et al., 2021). Unlike general-domain text, clinical narratives often express time in vague, relative, or context-dependent ways: for example, "a few days ago", "since surgery", or "post-op day 1" (Sun et al., 2013a; Gumiel et al., 2021). Such expressions typically require reference to implicit timelines, which are not always explicitly documented. Moreover, temporal relations between clinical events must often be inferred, rather than directly stated (Costa and Branco, 2013; Styler IV et al., 2014; Moharasan and Ho, 2019). A critical and under-discussed issue in this context is *missingness*. Temporal information may be incomplete, inconsistently recorded, or entirely absent. Clinical data vary in structure and detail across providers and institutions, leading to fragmented or ambiguous timeline information (Sun et al., 2013b; Weiskopf and Weng, 2013). This lack of standardization complicates model training and evaluation, especially for tasks like temporal relation extraction, event ordering, and patient timeline reconstruction (Johnson et al., 2016; Lim et al., 2021a).

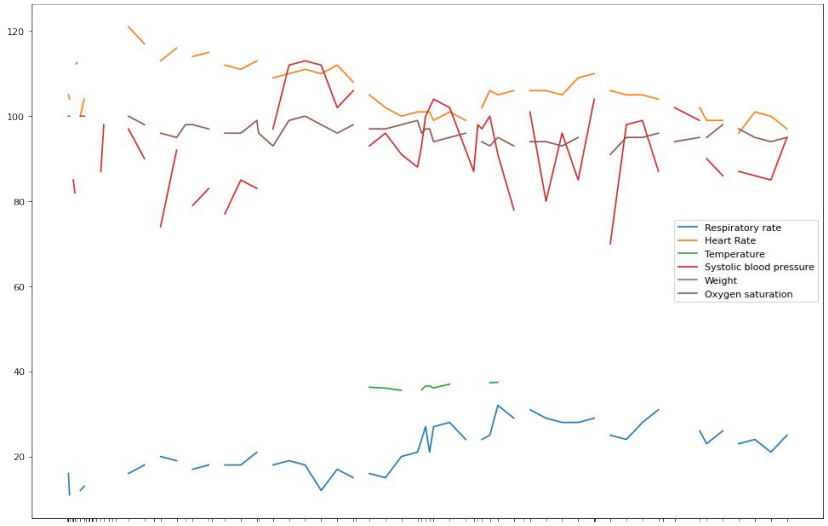



Figure 3.26: Depiction of temporal characteristics of medical data such as irregularity and missingness. Source : MIMIC-III

These complexities highlight the need for models that can robustly handle temporal information under uncertainty and variability. However, it remains unclear how effectively current medical language models manage these nuances, particularly when faced with implicit or incomplete time references. This leads us to ask:

**RQ 1.4**

What are the challenges medical models face on temporal-level?

### 3.5.1 Case Study VIII: Missingness

This chapter is based on work previously published in our article: *Agarwal, R., Sinha, A., Prasad, D. K., Clausel, M., Horsch, A., Constant, M., & Coubez, X. (2023). Modelling Irregularly Sampled Time Series Without Imputation. arXiv preprint arXiv:2309.08698.* Parts of the text, figures, and results are adapted from this publication.

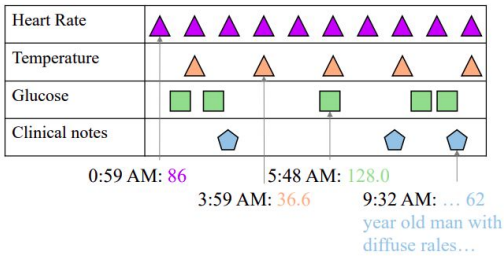



Figure 3.27: Illustration of *missingness* in Medical data (Zhang et al., 2023b).

Medical data, whether structured Electronic Health Records (EHRs) or free-text narratives, are often incomplete, irregular, and inconsistently recorded. This results in Irregularly Sampled Time Series (ISTS) multivariate time series where observations occur at non-uniform time intervals (see fig. 3.27). ISTS is not unique to medicine; it appears in domains such as meteorology (Mudelsee, 2002), seismology (Ravuri et al., 2021), and user behavior modeling (Zeng and Gao, 2022; Wu et al., 2013). However, in clinical settings, these irregularities are often driven by real-world constraints like medical necessity, provider decisions, and resource availability, making missingness an inherent and meaningful property of the data.

Despite its prevalence, missingness is typically treated as a nuisance something to be corrected through imputation, which fills in missing values using statistical or model-based estimates (Che et al., 2016; Shukla and Marlin, 2018; Lim et al., 2021b). This transforms ISTS into regularly sampled sequences (see Fig. 3.28a), enabling the use of standard time series models. However, imputation can introduce artifacts, distort temporal dynamics, and lead to distribution shifts that harm model generalization (Ipsen et al., 2020; Zhang et al., 2023c). Moreover, imputation assumes knowledge of the underlying missingness mechanism which is often latent, non-random, and domain-specific (Ma and Zhang, 2021; Rubin, 1976). In many cases, the pattern of what is missing and when may itself carry valuable clinical information. Yet, few existing models are designed to directly leverage this structure. While recent work has proposed non-imputation-based methods for handling ISTS (Horn et al., 2020; Vaswani et al., 2017), these approaches often fail to fully capture the temporal dependencies and semantics of missingness in medical data. As a result, it remains unclear how well current models perform when exposed to the types of irregularity and incompleteness commonly found in real-world healthcare settings.

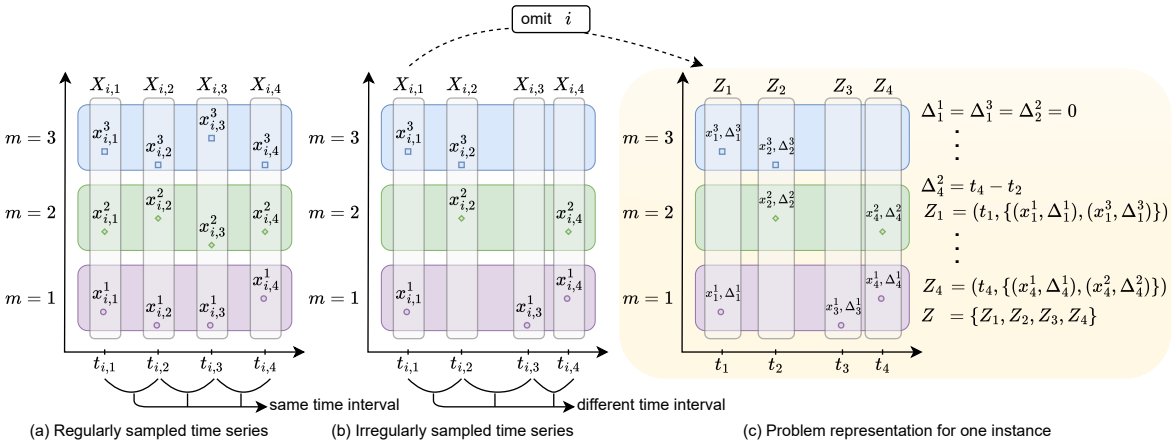
In this work, we investigate the following research question :

**Research Question 1.2e:** *What is a better way to learn from missingness found in medical data?*

Concretely, in the following study we propose a non-imputation model that takes as input completely raw data without any processing or treatment of missing values and dynamically adapts its architecture to learn from the pattern of missingness itself.

# Background

Irregularly Sampled Time Series (ISTS) follows the definition of Regularly Sampled Time Series (RSTS; introduced in Section 2.2.3), except not all sensors will be measured at each time step[29], leading to an irregular sampling of each sensor, and $t_{i,2} - t_{i,1}$ need not be equal to $t_{i,3} - t_{i,2}$. ISTS is mathematically given as $X_{i,j} \subseteq \{x^1_{i,j}, ..., x^s_{i,j}\}$, where $s$ is the total number of sensors. A snapshot of ISTS for $i^{th}$ instance is shown in Fig. 3.28b where $X_{i,1} = \{x^1_{i,1}, x^3_{i,1}\}$, $X_{i,2} = \{x^2_{i,2}, x^3_{i,2}\}$, $X_{i,3} = \{x^1_{i,3}\}$ and $X_{i,4} = \{x^1_{i,4}, x^2_{i,4}\}$. For simplicity, we omit the subscript $i$ representing an instance and consider only one




(a) Regularly sampled time series (b) Irregularly sampled time series (c) Problem representation for one instance

Figure 3.28: (a) A snapshot of multi-variate regularly sampled time series for $i^{th}$ instance. $m$ represents the index of the sensor. (b) A snapshot of multi-variate irregularly sampled time series (ISTS) for $i^{th}$ instance. (c) Problem representation of the ISTS with respect to one instance by omitting the subscript $i$. (Best viewed in color)

instance to discuss the problem and the working of the proposed model. In that sense, each measured value is given by $x^m_j$ (instead of $x^m_{i,j}$) and it is received at time $t_j$ (instead of $t_{i,j}$). Let us denote this instance by $Z = X_i$ and the set of values of measured sensors at time $t_j$ by $Z_j$. Based on this, at each time step $t_j$, we represent the measured sensors as $Z_j = (t_j, \bigcup_{\forall m \in \mathbb{A}_j} \{(x^m_j, \Delta^m_j)\})$ where $\mathbb{A}_j$ is the set of sensors measured at time $t_j$, given by $\mathbb{A}_j \subseteq \mathbb{M}$[30]. $\Delta^m_j$ denotes the time delay between two successive values measured by sensor $m$, *i.e.*, $\Delta^m_j = t_j - t_k$, where $k = max$[31]$(1, ..., j-1)$ such that $m \in \mathbb{A}_k$. Thus the whole input data is given by $Z = \bigcup_{j=1}^{l} \{Z_j\}$ where $l$ is the number of time steps. Following the previous paragraph, we visually present the data representation and the corresponding equations in Fig. 3.28c.

The motivation of Switch LSTM Aggregation Network (SLAN) is propelled by the effectiveness of the sequence model (LSTM) in handling time series data (Che et al., 2016). However, a single LSTM is incapable of modeling ISTS without imputation.

[29] We denote a timestep as $t_{i,j}$ where i is the index of the sensor and j is the index of the timestep.

[30] As previously mentioned in §2.2.3, $\mathbb{M}$ is the set of indices of sensors and can be written as $\mathbb{M} = \{1, \ldots, s\}$.

[31] $max$ operator is used to obtain the last index of time when for the sensor $m$ before $t_j$.

Therefore, we devise a strategy of employing one LSTM per sensor. Since ISTS has irregular sampling, we propose a simple switch layer that facilitates the activation of only those LSTMs whose corresponding sensors are measured. Furthermore, we introduce global and local summary states to share information between all sensors.

**Our Proposed Model: SLAN (Switch LSTM Aggregation Network).** SLAN is an adaptive LSTM-based model that dynamically changes its architecture depending on the measured sensors at any time point by utilizing a switch layer. The architecture of SLAN is presented in Fig. 3.29a. It consists of a pack of LSTMs such that there is a one-on-one connection between a sensor and an LSTM block. The switch layer facilitates this (see the yellow-colored box in Fig. 3.29a). Each sensor is connected to its corresponding LSTM block by a switch. A switch goes "on" if its corresponding sensor is measured; otherwise, it stays off. The "on" switch results in activating its corresponding LSTM block, thus, eliminating the need for any imputation. The LSTM block outputs a Long-Term Memory (LTM) and a Short-Term Memory (STM) (Fig. 3.29b). The LTM of each activated LSTM block is aggregated to produce a global summary state and passed on to all the LSTM blocks for the next time step as input. This aids the LSTM blocks with summarised information.

LSTM allows sequential processing of the time series, preserving their arrival order. However, still, it is necessary to model the time information associated with each input. This is more so in the case of ISTS since the time interval is not fixed. We draw upon the many methods presented in the literature to model time information and utilize Time2Vec (Kazemi et al., 2019) for its demonstrated effectiveness.

The previous short-term memory (STM) of each activated LSTM block is decayed based on the vector representation of time delay and decay function (discussed below). This decayed STM is passed as an input for the next measured time point. This acts as a local summary for each sensor. Finally, at the last time point, the STM(s) from each LSTM block is concatenated with the aggregated LTM as seen in the concat layer in Fig. 3.29a. The concat layer is then fully connected to a 2-node output layer for binary classification. The pseudo-code of SLAN and the unrolled SLAN architecture for the data example given in Fig. 3.28c is discussed in the *Algorithm* and *Unrolled SLAN* section in Supplementary, respectively.

SLAN consists of $s$ LSTM blocks $\{L^1, ..., L^s\}$ where $L^m$ is associated with sensor $m$. We define the switch layer ($\mathbb{S}_j$) as the set of switches kept "on" based on the measured sensors at time $t_j$. Since there is a one-on-one correspondence between a switch and its corresponding measured sensor, we borrow the representation of $\mathbb{S}_j$ as the indices of the sensors measured at time $t_j$ from section *Problem Formulation*, thus $\mathbb{S}_j = \mathbb{A}_j$. Note that the switch layer in SLAN is explicit information to the model compared to the implicit concatenation of the observation mask with the inputs in the case of the Transformer. A detailed discussion on the difference is presented in *Switch Layer in SLAN vs Observation Mask in Transformer* section in Supplementary. Each active LSTM block ($L^m$) at time $t_j$ takes the sensor value ($x_j^m$), STM ($h_{j-1}^m$), LTM ($c_{j-1}^m$) and

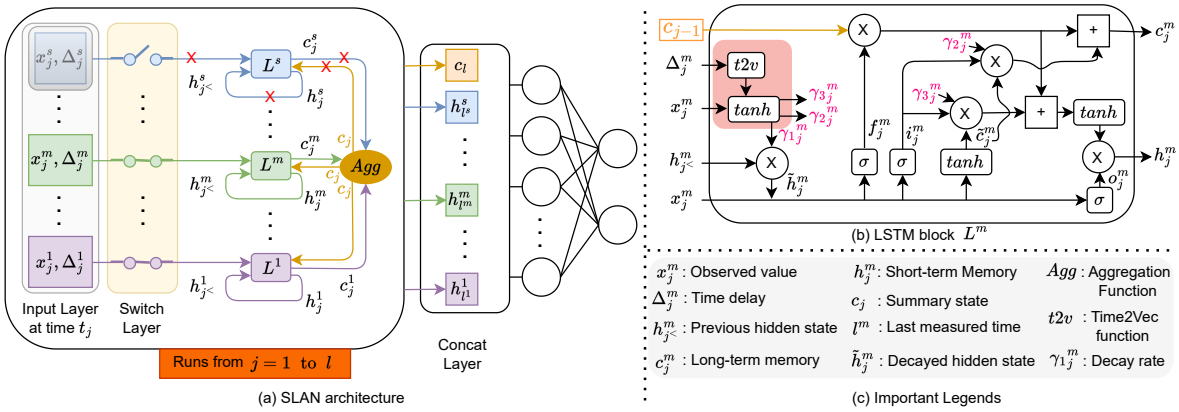



Figure 3.29: SLAN Architecture and its internal component illustration.

time delay ($\Delta_j^m$) as inputs and outputs $h_j^m$ and $c_j^m$, given by

$$(h_j^m, c_j^m) = L^m(x_j^m, h_{j-1}^m, c_{j-1}^m, \Delta_j^m) \quad \forall m \in \mathbb{S}_j \tag{3.5}$$

where $L^m \, \forall m \in \mathbb{A}_j$ are active based on $\mathbb{S}_j$ at time $t_j$. An aggregate function is employed on the LTM of the active LSTM blocks to get a summary state ($c_j$) at $t_j$. Any function that can group multiple values to give a single summary value can be used as an aggregation function and is represented by $agg()$. Some examples of aggregation functions are mean, max, and attention. The summary state is given as

$$c_j = agg(\bigcup_{\forall m \in \mathbb{S}_j} \{c_j^m\}) \tag{3.6}$$

The $c_j$ is used as an input for the next time step for every active LSTM block. An active LSTM at $t_j$ might not be active at $t_{j-1}$. Thus, the STM input to $L^m$ at $t_j$ is represented by $h_{j<}^m$ (instead of $h_{j-1}^m$) where $j^< = \max(1, ..., j-1)$ such that $m \in \mathbb{S}_j$. Therefore, eq. 3.5 can be updated as:

$$(h_j^m, c_j^m) = L^m(x_j^m, h_{j<}^m, c_{j-1}, \Delta_j^m) \quad \forall m \in \mathbb{S}_j \tag{3.7}$$

Finally, the hidden states (or STM) of all the LSTM blocks and the summary state are concatenated to give a final output. Note that all the sensors may not be observed in the last timestamp $t_l$. Therefore, we represent the last measure time for each sensor by $t_{l^m}$, such that $t_{l^m} \leq t_l$. Thus, the concat layer is given by $C = \{c_l, h_{l^1}^1, ..., h_{l^s}^s\}$. A fully connected network is employed to get a final prediction from $C$ as follows

$$\hat{y} = F(C) \tag{3.8}$$

Since the measurement of each sensor is irregular, we employ a time-decay function on the hidden states inspired by (Kazemi et al., 2019). The time-decay function ensures that the previous local summary is adjusted based on the time delay ($\Delta_j^m$) of each sensor. Since each sensor is different, the decaying function should differ for each

sensor. Thus, a trainable time-decay function is employed. The decay function is given as

$$\begin{aligned} {\gamma_1}_j^m = tanh(W_{\gamma_1}^m x_j^m + V_{\gamma_1}^m t2v(\Delta_j^m) + b_{\gamma_1}^m), \quad \text{where} \\ t2v(\Delta_j^m) = sin(\omega_j^m \Delta_j^m + \varphi_j^m) \end{aligned} \tag{3.9}$$

Here, $t2v$ is the Time2Vec function with $\omega_j^m$, and $\varphi_j^m$ as the learnable parameters. The sine function in Time2Vec helps capture periodic behaviors without the need for feature engineering. The $W_{\gamma_1}^m, V_{\gamma_1}^m$, and $b_{\gamma_1}^m$ are the parameters of the decay function. Consequently, the decay of the hidden state is given by

$$\tilde{h}_j^m = {\gamma_1}_j^m \odot h_{j<}^m \tag{3.10}$$

where $\odot$ is the element-wise dot product.

We employ TimeLSTM (Zhu et al., 2017) as an LSTM block in our study. The gates of $L^m$ at time $t_j$ are denoted by forget gate ($f_j^m$), input gate ($i_j^m$), output gate ($o_j^m$) and cell state ($\tilde{c}_j^m$). Based on the decayed hidden states ($\tilde{h}_j^m$) given by eq. 3.10 and summary state ($c_{j-1}$) given by eq. 3.6, the gates are determined as

$$\begin{aligned} f_j^m &= \sigma(W_f^m x_j^m + V_f^m \tilde{h}_j^m + b_f^m) \\ i_j^m &= \sigma(W_i^m x_j^m + V_i^m \tilde{h}_j^m + b_i^m) \\ o_j^m &= \sigma(W_o^m x_j^m + V_o^m \tilde{h}_j^m + b_o^m) \\ \tilde{c}_j^m &= tanh(W_c^m x_j^m + V_c^m \tilde{h}_j^m + b_c^m) \end{aligned} \tag{3.11}$$

The final short-term and long-term memory depends on the decayed cell state achieved via ${\gamma_2}_j^m$ and ${\gamma_3}_j^m$ (equation 3.9) and is given by

$$\begin{aligned} c_j^m &= f_j^m \odot c_{j-1} + i_j^m \odot \tilde{c}_j^m \odot {\gamma_2}_j^m \\ h_j^m &= o_j^m \odot tanh(f_j^m \odot c_{j-1} + i_j^m \odot \tilde{c}_j^m \odot {\gamma_3}_j^m) \end{aligned} \tag{3.12}$$

# Methodology

**Dataset.** We consider MIMIC-III (M-3), and Physionet 2012 (P-12) datasets to showcase the efficacy of SLAN. We prepare the datasets by following SeFT (Horn et al., 2020). Table 3.30 presents the details for both the datasets. #Instances is the number of patient records in the datasets, #Sensors is the number of features/sensors in each instance, # Static is the number of static variables, #Observations is the average number of observations recorded in each instance, i.e., the number of time steps, #Num-Imputation is the number of imputation or missing values and Imbalance is the percentage of instances with a minority class label.

| Dataset | MIMIC-III | Physionet 2012 |
|---|---|---|
| #Instances | 22110 | 11988 |
| #Sensors | 17 | 37 |
| #Static | 0 | 6 |
| #Observations(avg.) | 77.7 | 74.9 |
| #Num-Imputation | $1.8 \times 10^7$ | $2.8 \times 10^7$ |
| Imbalance (%) | 13.22 | 14.24 |

Table 3.30: ISTS Dataset Description.

**Baselines.** We consider both non-imputation and imputation baselines. Among imputation, GRU-D (Che et al., 2016), IP-Nets (Shukla and Marlin, 2018), and ViTST (Li et al., 2024) are considered, which are described below. The non-imputation baselines are Transformer (Vaswani et al., 2017), SeFT (Horn et al., 2020), Raindrop (Zhang et al., 2021), CoFormer (Wei et al., 2023), and IVP-VAE (Xiao et al., 2024), details of which can be found in Appendix appendix B.2.

1. GRU-D (Che et al., 2016) exploits missingness by considering two main missingness representation methods, masking and timestamps, to devise effective solutions to characterize the missing patterns. The proposed model aims to use the masking information and temporal pattern in the missingness via the two trainable decay terms. The decay is calculated as

$$\gamma_t = exp\{-max(0, W_\gamma \delta_t + b_\gamma)\} \tag{3.13}$$

where $\gamma$ is the decay parameter at time $t$, $W$ and $b$ are model parameters to learn the decay. GRU-D decays the hidden states as

$$h_{t-1} = \gamma_{h_t} \odot h_{t-1} \tag{3.14}$$

where $h_{t-1}$ is the hidden state from time $t-1$ and $\gamma_{h_t}$ is decay value of hidden state at time $t$. GRU-D further imputes the input missing value whenever the input data is missing. The following equation does the imputation

$$x_t^d = m_t^d x_t^d + (1 - m_t^d)\gamma_{x_t^d} x_{t'}^d + (1 - m_t^d)(1 - \gamma_{x_t^d})\tilde{x}^d \tag{3.15}$$

Here, $m_t^d$ represents the masking value which is 1 if the sensor is measured otherwise 0, $\gamma_{x_t^d}$ is the decay factor, $x_{t'}^d$ is the last observation of the $d^{\text{th}}$ variable ($t' < t$) and $\tilde{x}^d$ is the empirical mean of the $d^{\text{th}}$ variable. Thus, the missing input feature is imputed whenever not measured.

2. IP-Nets (Shukla and Marlin, 2018) or Interpolation-Prediction Networks consist of an interpolation network followed by a prediction network. IP-Nets convert ISTS to regularly sampled time series (RSTS) in the interpolation network. It uses the information from each time series to interpolate values of all the other

time series. IP-Nets considers a set of reference time points $r = [r_1, ..., r_T]$. All the reference time points are evenly spaced within its interval. For each sensor of an instance, IP-Nets output three interpolants (cross-channels, transient component, and intensity) corresponding to each reference point and a sensor. Thus, the interpolation network takes $i^{\text{th}}$ ISTS instance ($X_i$) as input and outputs $i^{\text{th}}$ RSTS interpolated output ($\hat{X}_i$) where the dimension of $\hat{X}_i$ is $(3s) \times T$. Here, $s$ is the number of sensors/features, $T$ is the number of reference time points, and 3 represents the number of interpolants corresponding to each time point for each sensor. Finally, in the prediction network, $\hat{X}_i$ is used as an input to produce the final prediction as $\hat{y}_i = g_\theta(\hat{X}_i)$.

3. ViTST (Li et al., 2024) or Vision Time Series Transformer converts each sample of ISTS data into line graphs. These graphs are subsequently organized into a standard RGB image format. The process involves plotting timestamps on the horizontal axis and observed values on the vertical axis of the line graph, with observations connected chronologically using linear interpolation to address missing values. Each sensor or feature generates a line graph that is arranged into a single image following a predefined layout. The vision transformer, specifically the Swin Transformer, is utilized for the classification of the created image. To integrate static features, ViTST transforms them into text using a template and encodes this text with a RoBERTa-base text encoder. The text and image embeddings are then concatenated to facilitate classification.

**Task.** For both datasets, the in-hospital mortality prediction task is considered which is a binary classification task. Given the multi-variate irregularly sampled time series data, the model has to predict 0 or 1 based on if or not the patient survives.

**Evaluation Metrics.** We use the Area Under the Receiver Operating Characteristic Curve (AUROC) and the Area Under the Precision-Recall Curve (AUPRC) as comparison metrics considering the imbalance in both the datasets (See table 3.30). The AUROC measures a model's ability to distinguish between positive and negative classes across all possible classification thresholds. It is computed from the Receiver Operating Characteristic (ROC) curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR). A higher AUROC indicates better ranking performance of positive instances over negative ones. In contrast, the Area Under the Precision-Recall Curve (AUPRC) plots precision versus recall at various thresholds and is particularly useful in imbalanced classification settings, where accurately identifying the minority (positive) class is critical. AUPRC provides a more informative evaluation in such cases by focusing on the quality of positive predictions.

**Implementation Details** We consider the train-val-test split of all datasets provided in SeFT. To handle the imbalance, we resort to a weighted oversampling strategy. Weighted oversampling involves preparing the training batch by sampling the data based on the class weights given by the inverse frequency of the class. The models are trained for 20 epochs with an early stopping of 5 on AUPRC to avoid overfitting. SLAN

uses cross-entropy loss, AdamW optimizer, data standardization, and mean aggregate function. The size of short-term and long-term memory size is 64, and the learning rate is 0.0005. The learning rate is adaptive with decay by a factor of 0.5 after each epoch without improvement. The batch size is 16, and the dimension of the time embedding vector is 16 for both datasets. Since the P-12 dataset has 6 static features (i.e., RecordID, Age, Gender, height, ICUType, and Weight), the embedding of these features is concatenated in the final concat layer before applying a fully connected layer for prediction. The size of the embedding is kept equal to the size of the global summary state. We ran experiments on an NVIDIA DGX A100 machine.

## Results & Discussion

The performance of SLAN on M-3 and P-12 datasets is presented in Table 3.31. SLAN outperforms all the baselines on both the datasets for both metrics. SLAN outperforms the second-best results by 1.2% and 8.9% in absolute AUPRC points, and 0.2% and 1% in absolute AUROC points for M-3 and P-12, respectively.

| Type | Model | MIMIC-III | | Physionet 2012 | |
|---|---|---|---|---|---|
| | | AUPRC | AUROC | AUPRC | AUROC |
| Imputation | GRU-D | 45.91 ± 1.34 | 83.43 ± 0.66 | 49.63 ± 1.17 | 84.94 ± 0.29 |
| | IPNets | 48.70 ± 0.67 | 84.90 ± 0.26 | 50.02 ± 0.61 | <u>85.54 ± 0.42</u> |
| | ViTST | 47.88 ± 0.49 | <u>85.49 ± 0.82</u> | 48.53 ± 1.05 | 84.27 ± 0.37 |
| Non-Imputation | Transformer | 48.88 ± 1.01 | 84.89 ± 0.53 | 49.37 ± 0.77 | 84.23 ± 0.14 |
| | SeFT | 46.01 ± 1.06 | 85.43 ± 0.26 | <u>50.69 ± 0.89</u> | 85.28 ± 0.28 |
| | Raindrop | 35.76 ± 0.29 | 77.18 ± 0.20 | 42.28 ± 1.48 | 79.34 ± 0.19 |
| | CoFormer | <u>50.51 ± 0.90</u> | 85.08 ± 0.56 | 48.67 ± 2.55 | 85.12 ± 0.96 |
| | IVP-VAE | 47.02 ± 0.75 | 84.80 ± 0.19 | 47.35 ± 0.72 | 85.12 ± 0.59 |
| | **SLAN** | **51.12 ± 0.57** | **85.63 ± 0.07** | **55.20 ± 0.65** | **86.42 ± 0.13** |

Table 3.31: Comparison of various methods on M-3 and P-12 datasets. The **best** and <u>2nd best</u> performance is represented by **bold** and <u>underline</u>, respectively. The metric is reported as the mean ± standard deviation of three runs with different seeds.

As evident from Table 3.31, IP-Nets perform the best among the imputation models, surpassing other models in both evaluated metrics on the P-12 and the AUPRC metric on the M-3 dataset. Consequently, we compare SLAN with IP-Nets to assess SLAN's robustness under an increased number of missing observations. Therefore, we randomly drop 25%, 50%, and 75% of observed data in both the M-3 and P-12 datasets. As demonstrated in Figure 3.32, SLAN consistently outperforms IP-Nets across all scenarios. SLAN achieves gains in absolute AUPRC points of 7.32%, 5.49%, and 6.85% in the M-3 dataset and 7.18%, 7.17%, and 12.63% in P-12 for the respective data drop of 25%, 50% and 75%. These results assert the superiority of SLAN even in conditions characterized by a substantial proportion of missing observations.

To determine the efficacy of SLAN, we compare it with imputed SLAN. We consider three types of imputation, namely, forward fill (ffill), mean, and interpolation where imputation is performed in the time series data to fill out the missing values via the last measured value, global mean, and linear interpolation, respectively. As evident from Table 3.32, SLAN (highlighted in green) outperforms mean and interpolation. SLAN further surpasses ffill in the P-12 dataset and in the AUROC metric of the M-3 dataset.

| | MIMIC-III | | Physionet 2012 | |
|---|---|---|---|---|
| | AUPRC | AUROC | AUPRC | AUROC |
| *Imputation* | | | | |
| ffill | **51.46±0.49** | 85.18±0.46 | 51.06±0.49 | 85.07±0.37 |
| mean | 48.73±0.79 | 84.30±0.36 | 51.65±0.73 | 85.28±0.32 |
| inter. | 49.44±0.26 | 84.96±0.31 | 50.75±0.07 | 84.88±0.34 |
| SLAN | 51.12±0.57 | **85.63±0.07** | **55.20±0.65** | **86.42±0.13** |
| *Aggregation Function* | | | | |
| Max | 49.24±0.88 | 85.40±0.29 | 54.36±0.89 | 85.95±0.19 |
| Att | 50.38±0.96 | 85.59±0.47 | **55.37±0.10** | **86.44±0.16** |
| Mean | **51.12±0.57** | **85.63±0.07** | 55.20±0.65 | 86.42±0.13 |
| *Concat* | | | | |
| Only G.S. | 46.61±0.83 | 84.63±0.27 | 48.81±1.84 | 83.04±0.45 |
| Only L.S. | 50.96±0.51 | **85.80±0.42** | 54.43±0.31 | 86.20±0.20 |
| G.S.+L.S. | **51.12±0.57** | 85.63±0.07 | **55.20±0.65** | **86.42±0.13** |

Table 3.32: Comparison of SLAN for different Imputation methods, aggregation functions and variants of concat layer. Att stands for attention. G.S. stands for global summary state and L.S. stands for local summary state. G.S. + L.S. is the default setting of SLAN.

We compare the performance of SLAN for mean, max, and simple attention (Bahdanau et al., 2014) as the aggregation function to calculate the global summary state. In attention, the normalized weightage of the LTM of each active LSTM block is determined using a single-layer feed-forward neural network, followed by the weighted average of LTMs to output the global summary state. In max, the element-wise max is performed over the candidate summary states. The comparison is reported in the middle part of Table 3.32. The performance of max is lower than both attention and mean because max may downplay the contribution of many LTMs by highlighting just one, thus becoming sensitive to outliers. Among mean and attention, attention performs the best in P-12, whereas mean gives better results in M-3. Overall, the mean performs well since attention outperforms the mean by a margin of only 0.17 (AUPRC) and 0.02 (AUROC) in P-12, whereas it underperforms the mean by a margin of 0.74 (AUPRC) and 0.04 (AUROC) in M-3.

SLAN's concat layer consists of a global summary state and the local summary state of each sensor. We remove the global summary state from the concat layer to check the informativeness of the local summary state (*Only L.S.* in Table 3.32).

When compared with the default setting of SLAN, *i.e. G.S. + L.S.*, *Only L.S.* is slightly poorer (max by ∼1.39%) and even surpasses in AUROC on M-3 by 0.2%. This is because the local summary state contains individual sensor information and is also aided by the global summary state at each time step. Only global summary states in the concat layer (*Only G.S.*) perform 11.58% and 4.07% poorer than *G.S. + L.S.* in terms of AUPRC and AUROC on P-12. *Only G.S.* performs 8.82% and 1.17% poorer than *G.S. + L.S.* in AUPRC

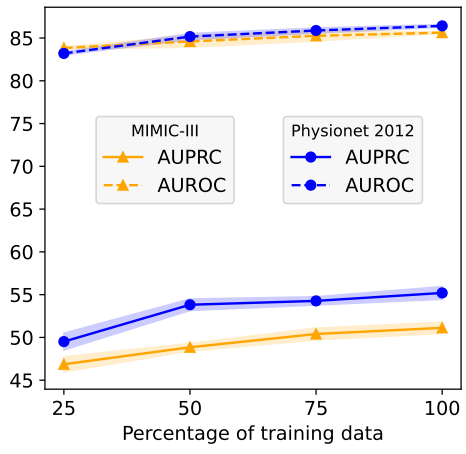



Figure 3.30: SLAN on different percentages of training datasets.

and AUROC, respectively for M-3. Indeed *Only G.S.* performs significantly poorer than *G.S. + L.S.*, still it contains sufficient summarised information to outperform baseline models like Raindrop in both datasets and GRU-D in M-3 (Table 3.31).

In the practical setting, it is important for any model to have data scalability, meaning the performance of the model on test data should improve as the amount of training data increases. We consider the first 25%, 50%, 75%, and 100% training data for both P-12 and M-3 and train our model on them. The average and the 95% confidence interval of 3 different runs of SLAN on the test data are shown in Figure 3.30. The performance of SLAN steadily increases with the increasing amount of training data on both datasets. The percentage improvement of AUPRC for M-3, when trained on 50% data compared to 25% data is 4.25%, 75% data compared to 50% data is 3.17%, and 100% data compared to 75% data is 1.43%. Transitioning from 25% to 50% data, we double the number of instances; thus, the percentage improvement is the highest. Whereas when trained on 100% data compared to 75% data, we add only 1/3rd data; thus, the percentage improvement is lowest. The same trend is followed in the AUROC of M-3, AUPRC and AUROC of P-12.

We use simple attention (Bahdanau et al., 2014) in the $Agg()$ unit to calculate the global summary state, which takes as input a set of LTMs $c_j^m$ from the activated $L^m$ at any time $t_j$. Our attention module contains a feed-forward neural network $\texttt{nn}(.)$ which calculates a set of scores for each $c_j^m$ as $\text{score}_j^m = \texttt{nn}(c_j^m)$ and are used to obtain the attention weights as $a_j^m = \text{score}_j^m / \sum_{k \in \mathbb{S}_j} \text{score}_j^k$. We explore $a_j^m$ as an attempt to interpret the characteristics of the information encoded in the global summary state. This ablation is motivated by the empirical evidence for global summary (only G.S. in Table 3.32), which performs better than some of the previous baselines. The sampling rate denotes the number of measurements per hour of a particular sensor. Ideally, a sensor with a high sampling rate should hold high importance since the model sees it most often. We consider the M-3 dataset for this ablation study. We sum the attention weights of each sensor across all the time steps, for each instance $i$. This is given by $\texttt{sumI}^m = \sum_{i=1}^{n_{test}} \sum_j a_{i,j}^m$, such that $m \in \mathbb{S}_j$ for $i^{\text{th}}$ instance and $n_{test}$ is the number of instances in the test dataset. Here $a_{i,j}^m$ represents the attention weight of $m^{\text{th}}$ sensor at time $t_j$ for $i^{\text{th}}$ instance. The summation weights of each sensor are then divided by the number of times each sensor is measured in the test dataset (denoted by $C^m$) as $\texttt{meanI}^m = \texttt{sumI}^m / C^m$. This is then

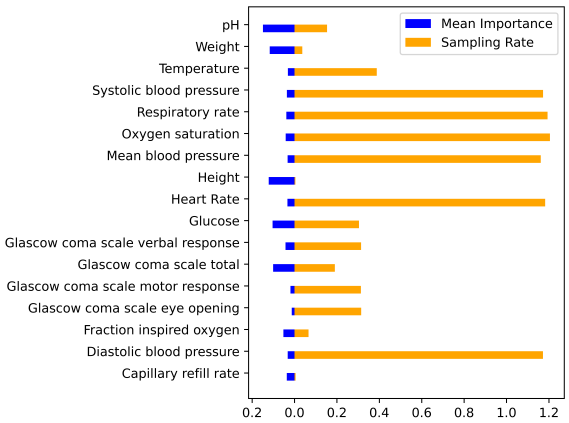



Figure 3.31: Comparison of the ranking of clinical variables *w.r.t.* sampling rate and mean importance.

normalized to get the importance of each sensor as $\texttt{normI}^m = \texttt{meanI}^m / \sum_{k=1}^{s} \texttt{meanI}^k$. We compare the mean importance ($\texttt{normI}^m$) and the sampling rate ($r^m$) of sensors in Figure 3.31. Sensors with higher sampling rates are oxygen saturation, respiratory rate, and heart rate. Whereas the most important sensors are pH, height, and weight. Even though pH has the fifth-lowest sampling rate, it is the most important sensor in providing inference. Thus, frequently measured sensors may not be the most important sensors.

**Summary** In this case study, we looked into the *missingness* phenomena via irregularly sampled time series view of medical data and explored the different ways to handle it. We propose a Switch LSTM Aggregate Network to handle multivariate ISTS data without any imputation. Overall, our proposed model SLAN demonstrates better efficacy in handling missing data in comparison various imputation and non imputation based baselines and across various ablation studies. We also establish the superiority of SLAN compared to the imputation model even when additional data is missing.

Moreover, the SLAN framework can be extended for modeling multi-modality data, eg. clinical notes (Zhang et al., 2023c) and also be leveraged for streaming data modeling in an online setting with time-variant dimensions (Agarwal et al., 2023a).

**Limitations** The main limitation of the study is its positioning with the more recent and sophisticated architectures for time series such as CHRONOS (Ansari et al., 2024) and other large language models based approaches (Zhang et al., 2024b).

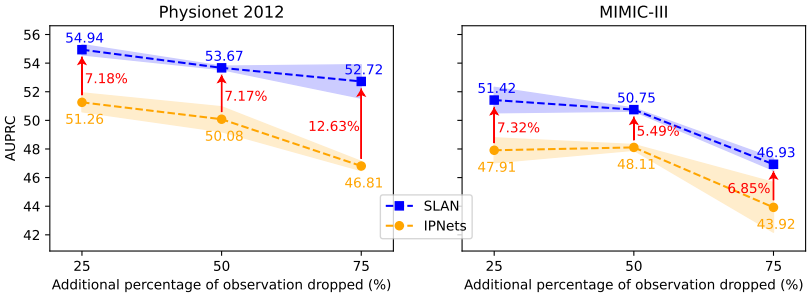



Figure 3.32: AUPRC of SLAN vs IPNets on P-12 and M-3 datasets with a drop of 25%, 50%, and 75% observed data. The red arrows show the % increase in the AUPRC of SLAN compared to IPNets with % increased value mentioned in the red-colored number. The blue and orange colored numbers represent the AUPRC of SLAN and IPNets, respectively.

## 3.6 Conclusion

This chapter explored the linguistic and structural limitations of single-sourced medical language models across multiple levels of linguistic complexity which includes token, sentence, document, and temporal level view and presented systematically case studies for them respectively. For each of these view, we evaluated medical models for their capability against different medical downstream tasks. We intentionally considered minimal setting without any added layer of sophistication in our experimental design such that the evaluations enables us to comment about the medical language models in an simple and interpretable manner.

We showed a common finding across the different case studies how restricted data sources constrain both the representational capacity and generalization behavior of medical models.

Token-level case studies revealed weaknesses in language models with respect to entity extraction, particularly when dealing with unseen noisy diseases and abbreviated medical names. Interaction with complex medical terminologies reaffirmed the limited behavior of general language models when exposed to medical domain.

Sentence-level analyses demonstrated how language models struggled with medical rich context with even minor contextual or stylistic variations leading to destabilizing predictions, reflecting an insufficient grounding in pragmatic cues. At the document level, while general domain language models struggled to capture relevant information for document understanding although, it still validates the hypothesis that integration of more information in form of more text and potential other knowledge sources can lead to improve its efficacy. However, this highlights simultaneously the fact for medical domain where data scarcity is common phenomena may lead to model's generalization capability if the dependence remain restricted to text data.

Finally, temporal view case study confirmed that, same as general domain practices, for medical domain it is not required to perform imputation for handling missingness. This positively confirms that it is possible to work out a way the missingness problem which could also potentially be translated for dealing with data scarcity problem as an alternative to data augmentation techniques.

Together, these findings suggest that the deficiencies of single-sourced models are not purely model-related but fundamentally data-centric. Errors originating at lower linguistic levels can lead to amplifying discrepancies between medical model and expected expert level deduction. These observations lay the groundwork for the next chapter, which examines whether multi-sourced modeling approaches can better capture the complexity inherent in medical language.

# Chapter 4

# Multi-source Medical Models



This chapter presents an investigation of multi-sourced medical models, exploring how language models can be enhanced through integration with external source of information. It begins with an overview of multi-sourced models (§4.1), followed by an in-depth examination of medical models augmented with external knowledge bases or knowledge graphs (§4.2). This includes case studies on entity extraction, medical paraphrasing, and document classification using external structured sources (§§4.2.1–4.2.3). The chapter then explores the use of additional modalities, such as multimodal learning with clinical text (§4.3, 4.3), and concludes by analyzing the role of large language models (LLMs) as assistants (§4.4), with a focus on their ability to generalize across writing styles in clinical data (§4.4.1). Overall, the chapter evaluates the extent to which multi-sourced approaches can overcome the limitations observed in single-sourced models by integrating diverse sources of knowledge and information in medical language processing.

## 4.1 What are Multi-Sourced Medical Models?

Intuitively, as the name suggests "Multi-Sourced" medical models are the models that uses more than one source of information to process medical data for any downstream task. This could involve usage of multiple sources of information as a joint training or utilization of multiple sources in a disjoint manner where the information is used sequentially. What we are interested in, is the integration of additional data source in order to build medical models to potentially overcoming the challenges that were identified in the previous section with single-sourced medical models.

Additional external knowledge source integration strategies may vary from case to case e.g., from the use of rule-based post-processing to knowledge graph embeddings, to hybrid architectures combining neural and symbolic components. However, incorporating external knowledge source also presents challenges, including alignment between multiple modalities or structured concepts, handling of conflicting sources, and scalability across large ontologies.

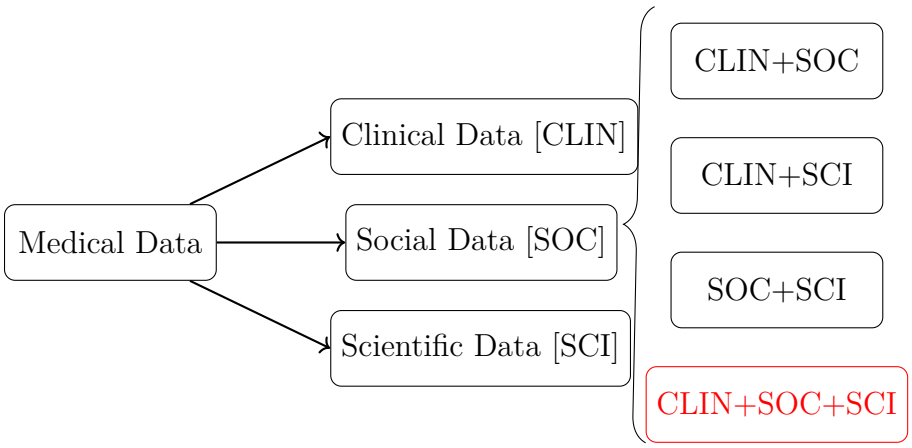



Figure 4.1: Different possibilities of Multi-source Models.

Figure 4.1 illustrates how individual data sources [CLIN], [SOC], and [SCI] can be combined in various configurations. The combination highlighted in red falls outside the scope of this thesis, as it does not involve a direct mapping between all three types of medical data. In practice, integrating additional sources into a medical model often involves augmenting a single-sourced model such as a pretrained language model (PLM) or a deep learning model with external resources like another PLM, a large language model (LLM), a knowledge base, or a knowledge graph. In this chapter, we present a series of case studies that explore the integration of such complementary sources. Our goal is to examine how multi-sourced medical models can be developed as a viable alternative to single-sourced models, and to address the following question:

**RQ 2**

Can Multi-Sourced Medical Models be more effective than Single-Sourced Medical Models when dealing with medical data ?

## 4.2 Use of External Knowledge Base or Knowledge Graph

Despite the promising capability of language models in medical NLP, many of the models remain fundamentally limited due to their reliance on a single source of information. While such models can capture rich contextual patterns, they often lack explicit domain knowledge required to make domain-specific sound inferences, particularly in edge cases, ambiguous phrasing, or low-resource settings.

To mitigate these limitations, any additional source of information, such as external knowledge bases (eg. biomedical ontologies) or a clinical modality (eg. clinical notes or x-rays) among many more, can offer a complementary source of structured, curated information. Resources such as SNOMED CT, UMLS, RxNorm, and International Classification of Diseases (ICD) ontologies provide standardized definitions, concept hierarchies, synonym mappings, and semantic relations between medical entities. Integrating such resources can enhance model performance by grounding predictions in domain knowledge, improving generalization, and offering interpretability.

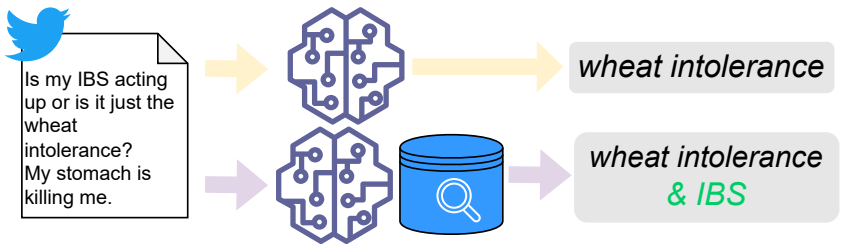



Figure 4.2: Illustration of External Knowledge Base for medical downstream task.

For example, fig. 4.2 shows a single-sourced model (on top), on the basis of its training data, only marks *wheat intolerance* as a disease mention. However, when given access to knowledge base, it is able to identify additionally *IBS* which is an abbreviation for "Irritable bowel syndrome". This illustrates that the use of external knowledge is particularly valuable in tasks such as entity normalization, filling out missing information, clinical grounding or clarification, where purely data-driven approaches may struggle with lexical variability, rare concepts, or contextual ambiguity.

In the following case studies, we explore the following research question:

**RQ 2.1**

Can the use of external knowledge source (eg. knowledge base or knowledge graph) help to mitigate the challenges faced by single-sourced models?

### 4.2.1 Case Study IX: External KB for Disease Identification

This chapter is based on work previously published in our article: *Sinha, A., Holgado, C. G., Clausel, M., & Constant, M. (2022). IAI@ SocialDisNER: Catch me if you can! Capturing complex disease mentions in tweets. In Mining for Health Applications, Workshop & Shared Task (# SMM4H 2022) (p. 85).* Parts of the text, figures, and results are adapted from this publication.

Traditional medical entity extraction models including those based on language models typically rely on the language representation capability present in a single source of data, such as formal clinical narratives or user-generated content (Li et al., 2019a; Kreimeyer et al., 2017). While these approaches found to perform reasonably well on clean, structured text, their ability to generalize often deteriorates when applied to social media originated medical data [SOC], which is inherently noisy, informal, and linguistically diverse (Limsopatham and Collier, 2016). The lexical and semantic variability found in [SOC] content introduces significant challenges for consistent entity identification, especially for disease-related terms expressed in layperson vocabulary often mixed with colloquial slang and non-standard way of using language (Tutubalina et al., 2018).

To address these challenges, we explore a multi-source learning paradigm by incorporating *external knowledge base*, in particular a medical knowledge base, as an additional layer of structured domain-specific information. We frame our exploration in the form of the following question:

**Research Question 2.1a:** *Can we improve LMs generalization for disease identification in [SOC] medical data with an external knowledge base?*

Resources like the Unified Medical Language System (UMLS) (Bodenreider, 2004), SNOMED CT, and MeSH offer curated taxonomies, synonym lists, and conceptual hierarchies that can enrich text-based models with clinical grounding. By aligning noisy free-form text with these structured ontologies, medical models can benefit from complementary cues — such as synonym resolution and concept disambiguation — that are not easily learnable from raw text alone. Prior work has shown that knowledge-informed modeling can improve extraction performance in biomedical contexts (Choi et al., 2016), particularly when domain-specific signal is weak or noisy. More concretely, we investigate expert curated knowledge base-assisted disease identification pipelines within this multi-source framework, aiming to enhance robustness and semantic fidelity in real-world scenarios.

## Methodology

**Task Description.** We are interested in the task of disease identification in tweets which was introduced in §3.2.1. The task involves, a social media post, (eg. tweet) that is given to the model as an input, and the model has to detect the span of the disease in the text. This was already investigated on the collection of language (embedding)

models. For this case study, we focus on the use of external knowledge base to see the impact on the capability of multi-sourced model setup against standalone language embedding models (listed in Table 4.2).

**Dataset.** We again utilize the SocialDisNER dataset (Sánchez et al., 2022) that was used in Section §3.2.1 for disease identification in tweets. The dataset comprises of a corpus of Spanish Twitter posts annotated for disease mentions. Table 4.1 shows examples of annotations from the corpora. The disease mention in the example are highlighted in blue font in order to show the non standard spelling and noise that surround them.

| ID | Input | Label | Start | End | Span |
|---|---|---|---|---|---|
| 25 | Si todavía no has leído nuestro último artículo en profundidad "Cómo me afecta la pandemia y el confinamiento en el control de mi #diabetes" te invitamos a conocerlo en https://t.co/xu27DC25n3 #DM2 #diabetESP | ENFERMEDAD (en translation: DISEASE) | 131 | 139 | diabetes |
| | | ENFERMEDAD | 198 | 201 | DM2 |

Table 4.1: Example from SocialDisNER dataset.

**Baseline Models.** We refer to the same collection of static and contextual language embeddings models listed in Table 4.2 that were used for the task in §3.2.1. For the static embeddings, the FLAIR Simple (Flair-S) package (Akbik et al., 2019), including classical es, en configuration[1], and a combined es+en+clinical configuration[2] was used. And for the contextual embeddings, the Flair-Transformer (Flair-T) package was used. This provides the option of integrating any pre-trained language models available on HuggingFace.

| Type | en | es | multilingual | domain |
|---|---|---|---|---|
| **STATIC** | / | ES | / | ES+CLIN<br>ES+EN+CLIN |
| CONTEXTUAL | BBUCN (Rawal, 2021) | BSCFN (Cañete et al., 2020) | XRL (Conneau et al., 2019)<br>BBMCN (Adelani, 2021)<br>WMN (Tedeschi et al., 2021) | XLRSC (Lange et al., 2021)<br>RBBCE (Carrino et al., 2021)<br>SDF (Chizhikova et al., 2022) |

Table 4.2: List of Language Embedding Models and their categorization.

**KB Integrated Pipeline.** Figure 4.3 shows the Illustration of the entire pipeline that is resulted by the integration of the knowledge base shown as the rightmost three-layered module. This module includes three noise-treatment parts:

[1]https://flairnlp.github.io/docs/tutorial-embeddings/classic-word-embeddings
[2]https://flairnlp.github.io/docs/tutorial-embeddings/flair-embeddings

1. *Disease list*: string-matching of disease mentions using an external custom list of disease words from a combination of the training disease mentions and online medical disease glossaries. This step facilitates the removal of agglutinated words or attached noise (e.g. *#labioRojoContraLaMigraña*⋆ [0-26 → 18-25]). The disease glossary was prepared manually by combining scraped disease list from different online sources leading to ∼130k disease list in Spanish.

2. *Glued words removal*: removal of outer noise specific to Twitter hashtags when no disease mention is matched. The beginning and the end of the mention are checked against the external KB to remove this specific noise (e.g. *#DíaNacionalDelPárkinson* [0-24 → 16-24], *#TodasContraElCáncerDeMama* [0-26 → 15-26]).

3. *Special characters (Spchar) removal*: removal of emojis, punctuation signs, and other related characters when no disease mention is matched from the list (e.g. *#autismo*♡ [0-9 → 1-8]).

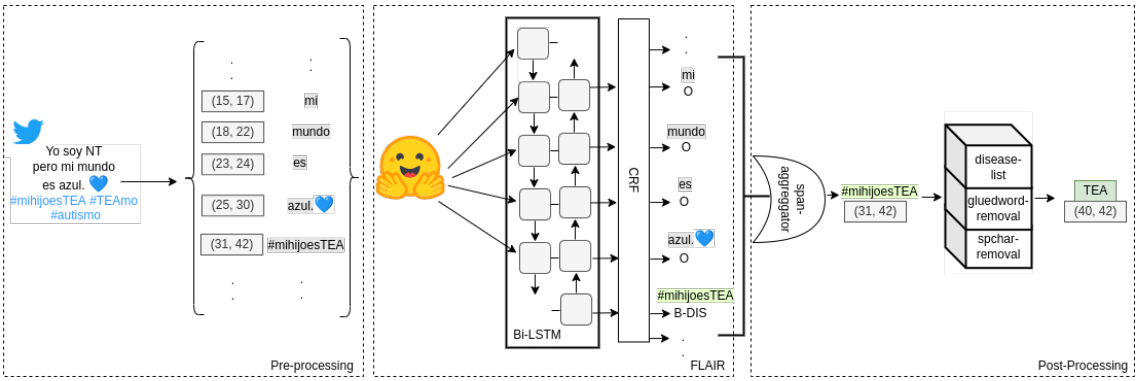



Figure 4.3: Complete Pipeline for Multi-sourced model based Disease Identification.

**Experimental Setup.** Firstly, using the base model (i.e. pre-processing and FLAIR module), predictions were done over the tokenized tweets considering the BIO-Tags. These predictions were then converted into the span format where every detected disease mention that was identified. In a last step, using the knowledge base, the aggregated original predicted spans undergo treatment. This fix of the original spans was performed on every detected disease mention when this contained any type of surrounding noise in the begin and/or end (See fig. 4.3).

**Evaluation.** We use F1-Score for Bio-tags evaluation and Strict-F1 and Lenient-F1 score for evaluating the performance of multi-sourced based language models. However, our this set of experiments the Bio-tagging and Lenient-F1 Scores remain relatively unchanged, and therefore, for the discussion of the results, we focus on Strict-F1 scores. This is because the integration of knowledge base is at the end of the base model i.e. its impact remains independent with respect to the type of embedding model.

## Results & Discussion

We present detailed results for single-sourced baseline setup in Table 4.3 corresponding to each language embedding model. For the baselines models, we notice that while they were efficient at identifying the disease mentions on a word-level (as indicated by Tag-F1 and Lenient-f1 scores in Figure 4.4), the presence of noise in the extracted disease mentions lead to low Strict-F1 scores.

| | LMs | Tag-F1 | Lenient-f1 | Baseline (*Strict-f1*) |
|---|---|---|---|---|
| en | **bbucn** | 0.87 | 0.914 | 0.298 |
| es | **es**† | 0.80 | 0.830 | 0.294 |
| multilingual | **bscfn** | 0.90 | 0.922 | 0.316 |
| | **xrl** | 0.93 | 0.950 | 0.317 |
| | **bbmcn** | 0.90 | 0.927 | 0.311 |
| | **wmn** | 0.89 | 0.925 | 0.306 |
| domain | **es+clin**† | 0.87 | 0.907 | 0.308 |
| | **es+en+clin**† | 0.88 | 0.910 | 0.313 |
| | **sdf** | 0.93 | 0.949 | 0.318 |
| | **xrlsc** | 0.93 | 0.955 | 0.319 |
| | **rbbce** | 0.93 | 0.951 | 0.318 |

Table 4.3: Social DisNER Final experiments results on Validation set. Baseline refers to Single-Sourced approach in Section 3.2.1.

Figure 4.4 presents the impact of knowledge base integration (added in the pipeline as the post-processing module) on the final predictions obtained from the model. The strategy of noise treatment[3] significantly improves the Strict-F1 scores by a range of 30 to 40% for all the embedding model.

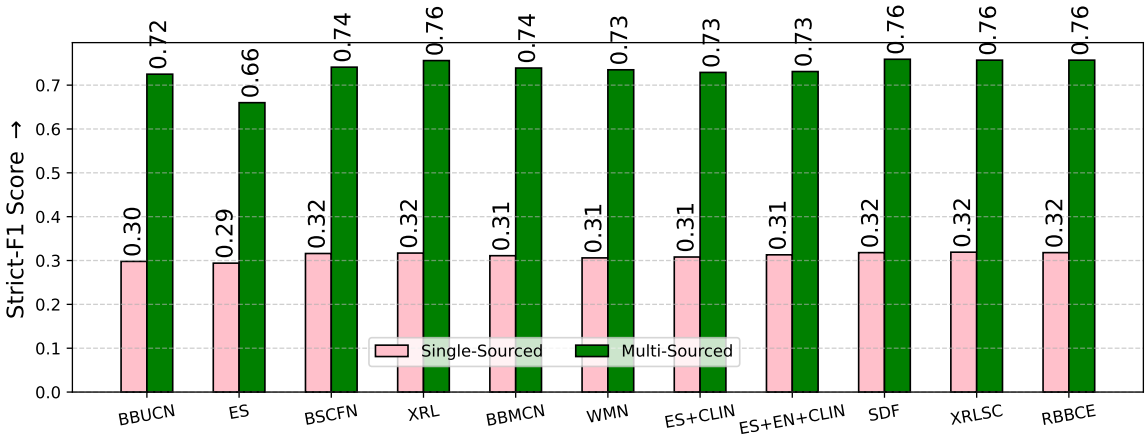



Figure 4.4: Effect of post-processing (PP) with metric Strict-F1.

During error analysis, we found that *complex and discontinuous named entities* were particularly problematic, as they often involved multiple disease mention identifications. Variable context boundaries (e.g., *#dolorneuropatico en #COVID19?* →

[3]We use blue font color to distinguish between target entities and noise.

single or multiple disease mentions?) was among another potential reason that may have lead to inconsistencies in span computation. We found that the error rate[4] for capturing long and discontinuous entities was 26.11% lower for Flair-T models than for Flair-S models. For noisy and agglutinated words such as *#HablemosDeVIH* or *#diabetestipo2*, Flair-T models performed more effectively than Flair-S models. In both cases Flair-T outperforms Flair-S, which likely due to subword tokenization that enable Flair-T to generalize better on unseen disease mentions. Despite the relatively small number of errors for Flair-T in this regard, we observed a negative impact of agglutinated words combined with emojis (e.g., *#autismo*♡).

Additionally, considering the character limit of tweets and the length of certain entities (e.g. *Enfermedad pulmonar intersticial difusa*), we encounter an increased use of acronyms (e.g. *#EPID*), which can be challenging for NER systems. We observed Flair-T models were 49.2% less prone to fail in detecting diseases' acronyms. Other transformations include flexions or verbal derivations (e.g. *resfriado→resfriarse*), where Flair-T was found to be consistently more effective. Furthermore, domain- and multilingual-specific embeddings have a comparable performance (refer Figure 4.4). Both performed better than es-models as they benefited from the knowledge from the clinical datasets (Lange et al., 2021) that were used to pre-train them. Irrespective of the adaptive fine-tuning, en-specific models performed lower than es-specific models. Their marginal difference can be attributed to common standard disease names used by users on Twitter.

Finally, we also found that domain-specific embedding models were able to identify previously unknown diseases as well as incomplete disease spans with respect to the gold annotations. There were instances where the identified disease mentions, with respect to the gold annotations, did not strictly correspond to actual diseases (e.g., *esquizofrenia cultural*). Moreover, new disease mentions were detected (e.g., *Alteraciones cutáneas*), in both Spanish and English (e.g., *#epilepsywarrior*), and complete spans were captured when the gold annotations were incomplete (e.g., *esteatosis hepática grado II–III*).

## Summary

To conclude and address the research question posed in this case study, we revisited the task of disease extraction on [SOC] data and demonstrated that integrating an external knowledge base into the disease extraction language model can significantly improve performance. Additionally, we observed that such an approach holds promise for enhancing the quality of gold-standard annotations.

These findings suggest that, while knowledge base integration is beneficial, it still lacks the generalization capability needed to robustly handle noisy, informal language in social media. Achieving that level of robustness will require deeper semantic modeling and a better understanding of the linguistic patterns present in user-generated content.

[4]This percentage was calculated len(number of diseases error + than 3 words)*100/len(number of diseases + than 3 worse on mentions in the validation set.)

**Limitations** a notable gap remains between the Lenient-F1 score (see fig. 4.4) and the improved Strict-F1 scores. This discrepancy highlights a limitation in the robustness of the knowledge-base-augmented model when handling unseen disease mentions. The use of fuzzy string matching to remove noise around target entities is constrained to known diseases present in the provided list. For previously unseen disease mentions, especially in noisy user-generated text, this becomes a challenging task due to the variability and agglutination of surrounding noise words (e.g., #ElVPHEsCosaDeTodos, #vacunavph, #DiaInternacionalContraelVPH).

### 4.2.2 Case Study X: External KB for Medical Paraphrasing

This chapter is based on work previously published in our article: *Buhnila, I., Sinha, A., & Constant, M. (2024, August). Retrieve, Generate, Evaluate: A Case Study for Medical Paraphrases Generation with Small Language Models. In Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024), pages 189–203, Bangkok, Thailand. Association for Computational Linguistics.* Parts of the text, figures, and results are adapted from this publication.

Traditional medical paraphrasing models, including those based on language models, typically rely on the language representation capability present in a single source of data, such as biomedical literature. While these approaches have been found to perform reasonably well within their training domains, their ability to generate diverse and accurate paraphrases often remains limited when dealing with the rich terminological landscape of medical and clinical jargon. The lexical and semantic complexity inherent in medical terminology characterized by technical terms, latin-derived nomenclature, specialized abbreviations, and domain-specific concepts introduces significant challenges for generating semantically equivalent yet linguistically varied expressions while maintaining clinical precision and accuracy.

| | Term | Paraphase |
|---|---|---|
| fr | myasthénie grave | est un trouble qui entraîne une faiblesse musculaire et une fatigue musculaire excessive |
| en | myasthenia gravis | is a condition that leads to muscle weakness and excessive muscle fatigue |

Table 4.4: Samples of medical term paraphrase from RefoMED dataset.

The nature of medical terminology intuitively suggests that effective paraphrasing may benefit from structured medical knowledge that provides relevant and factual information, rather than relying solely on surface-level linguistic patterns learned from text corpora. Recent advances in knowledge-enhanced generation have demonstrated that incorporating external knowledge bases through retrieval-augmented approaches (Lewis et al., 2020b) can provide the necessary grounding for accurate medical text generation, enabling models to leverage structured medical ontologies and relational knowledge to guide generation processes.

In this case study, we investigate whether incorporating an external medical knowledge base can address the limitations of single-source language models and help in improve medical paraphrasing .

Research Question 2.1b: *Can we improve medical LMs' generalization for simple definition generation for [SCI] medical terminologies with external knowledge base?*

More concretely, we examine how structured medical knowledge including terminological hierarchies, semantic relationships, and concept definitions curated from Wikipedia can guide paraphrase generation to produce diverse linguistic expressions while maintaining strict semantic equivalence and clinical accuracy across the complex

landscape of medical and clinical terminology.

## Methodology

**Task Description.** Our goal is to improve the quality of the output and obtain a short paraphrase of the term and not a complete description of the term (as shown in Table 4.4). Thus, we use prompt tuning techniques. However, we need curated sub-sentential paraphrase datasets in medical French for this task. Most of the paraphrase datasets contain only sentential paraphrases from general language in English: MSRP (Dolan et al., 2004), PPDB (Ganitkevitch and Callison-Burch, 2014), PAWS (Zhang et al., 2019b) or multilingual: TaPaCo (Scherrer, 2020) or ParaCotta (Aji et al., 2022).

| SLM | FR Encoder-Decoder | |
|---|---|---|
| **BaseSLM** | BARThez-orangesum-abstract<br>BioMistral-7B-SLERP-GPTQ | |
| **pRAGe** | **FR Encoder** | **FR Decoder** |
| **BioMistral** | DrBERT | BioMistral-7B-SLERP-GPTQ |
| | sent-CamemBERT | BioMistral-7B-SLERP-GPTQ |
| **BARThez** | DrBERT | BARThez-orangesum-abstract |
| | sent-CamemBERT | BARThez-orangesum-abstract |

Table 4.5: Configurations of different French encoders and decoders.

**Baselines.** We used BARTHEZ-OrangeSum-abstract (BARTHEZ[5]) (Eddine et al., 2020), a French seq2seq model, and BioMistral-7B-SLERP-GPTQ[6] (Labrak et al., 2024), a 4-bit precision Quantized Generative Pretrained Transformer (GPTQ) quantized (Frantar et al., 2022) multilingual medical model for training and inference efficiency[7]. We chose the BioMistral-7B-SLERP model (Shoemake, 1985) as it gave the best benchmark results on French datasets, according to the authors of the model (Labrak et al., 2024). We also tested the impact of finetuning the models for all configurations, using an existing sub-sentential paraphrase dataset in medical French, RefoMed (Buhnila, 2023). For finetuning, we used the Q-LoRA method (Dettmers et al., 2024), a computationally efficient finetuning method that reduces the number of parameters for BioMistral from 7B to 1.38B parameters.

**Our proposal: pRAGe.** We illustrate our proposed method in fig. 4.6. pRAGe is built on an encoder-retriever-decoder framework. We designed Pipeline for Retrieval Augmented Generation and Evaluation (pRAGe) to embed a medical query and to generate an output in a style that *translates* medical knowledge for patients in a simpler

[5]For readability reasons, we will hence refer to BARTHEZ-OrangeSum-abstract as BARTHEZ.

[6]BioMistral was pre-trained on 3 bilion tokens data from PubMed Central from Mistral. Less than 1,25% of the data is a GPT-3.5 Turbo automatic translation in French and other 8 languages.

[7]https://huggingface.co/LoneStriker/BioMistral-7B-SLERP-GPTQ

language, e.g. *rhizarthrose* → *arthrose du pouce* (rhizarthrosis → arthrosis of the thumb). Therefore, we paired models for the general language with medical models. We tested different configurations of non-proprietary encoders and decoders, as shown in Table 4.5. We used the general French encoder model sent-CamemBERT ((Reimers and Gurevych, 2019; Martin et al., 2020)) and the domain-specific model DrBERT (Labrak et al., 2023a), a French BERT type model for the medical field. The pRAGe pipeline encodes in embeddings the input query and the Wikipedia knowledge base, RefoMed-KB (introduced below this paragraph). We performed prompt engineering to guide the decoder towards the expected output in both experimental settings, base model inference and pRAGe pipeline. The task attributed to the model is "to answer the user's question with a paraphrase, explanation or short definition" (See prompt in Figure 4.5).

**fr** Expliquez-moi le terme médical en mots simples, par une paraphrase ou une courte définition :

**en** Explain the medical term to me in simple words, through a paraphrase or a short definition :

Figure 4.5: Our prompt template for inference.

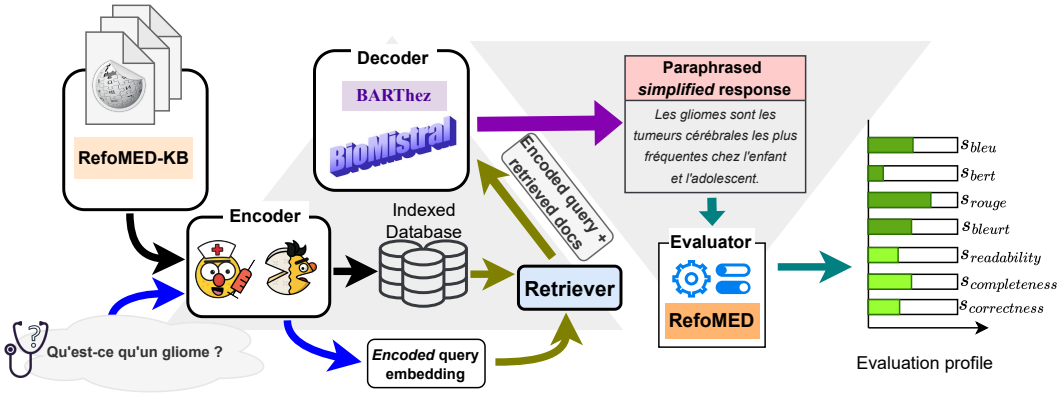



Figure 4.6: Illustration of pRAGe experimental pipeline.

**Dataset.** We used the same open-source dataset of medical sub-sentential paraphrases in French, RefoMed[8] (Buhnila, 2023) that was introduced in Section 3.2.3 as input queries and to finetune the models in pRAGe. The RefoMed corpus is made of 6,297 pairs of unique medical terms and their corresponding sub-sentential paraphrases in French. We split the RefoMed dataset for finetuning by unique term entry while staying in the range of the classic 60-20-20 train-validation-test split proportion. The validation and test sets were used to build the knowledge base for the RAG system, RefoMed-

[8]https://github.com/ibuhnila/refomed

KB. The resulting split is as follows: 3,981 term-paraphrase pairs for training, 1,063 for validation and 1,253 pairs for testing.

**External Knowledge Base: RefoMED-KB.** The next step was to build the Knowledge Base (KB) for the medical terms from the validation and testing sets. We automatically extracted top-3 Wikipedia articles where the terms appear in the title of the article using the Python *wikipedia* library. We extracted the first 20 lines of each relevant wikipedia page and we obtain a medical knowledge base in French of 20,402 sentences (1,708,034 tokens) about the 1,253 medical terms from the test list.

**Evaluation metrics.** Evaluations for the complete RAG framework can be divided into two categories: intrinsic and extrinsic. For extrinsic evaluation, we check for hallucination by evaluating the rate of medical correct answers (Huang et al., 2023). For intrinsic evaluation, we perform manual evaluation by checking the quality of responses generated.

Several metrics are used in the context of State-of-the-art (SOTA) research on text generation evaluation: ROUGE (Lin, 2004), calculates the n-grams overlap (recall), BLEU (Papineni et al., 2002), computes the number of similar n-grams between the output and the reference (precision), BERTscore (Zhang et al., 2019a), compares the embeddings of tokens that match in the output and reference text, while BLEURT (Sellam et al., 2020), computes the semantic similarity and lexical difference between them. MEDCON (Yim et al., 2023) is a metric that computes the F1-score of the UMLS concepts found both in the output and the reference text (however, available only for English). For this case study, we evaluate the similarity between the generated output text and the reference text in French. We use the following evaluation metrics: `bleu`, `rouge`, `bleurt`, and `bertscore`.

1. **Automatic Evaluation** We define a score metric for evaluating the generated response set from pRAGe. For any $i_{th}$ query, if $p_i$ is the generated response from the RAG pipeline and $\mathcal{R}_i$ is the list of reference paraphrases:

$$S_\Omega = \frac{\sum_{i=1}^{N} \texttt{max}(\{\Omega(p_i, r_{ij}) \forall r_{ij} \in \mathcal{R}_i\})}{N} \tag{4.1}$$

   where $N$ is number of queries and $\Omega$ is a lexical or semantic similarity comparison metric such as `bleu`, `rouge`, `bleurt`, `bertscore`, etc.

2. **Human Evaluation** We conduct a fine-grained evaluation of the generated paraphrases. A set of 1200 examples [9] were manually analyzed by 3 French proficient linguist annotators with the criteria described in Table 4.6.

[9] 50 examples from 24 different configurations where we have 6 unique configs, as presented in Table 4.5, with 2 fine-tuning varians and 2 token-limit variants.

| Criterion | Scale | Description |
|---|---|---|
| ROUGE (↑) | 0-1 | Measures n-gram overlap with the reference text, emphasizing recall. |
| BLEU (↑) | 0-1 | Measures n-gram precision between the candidate and reference texts. |
| BERTScore (↑) | 0-1 | Assesses semantic similarity using contextual embeddings (token-level cosine similarity). |
| BLEURT (↑) | -1 to 1 | Uses a learned model to estimate semantic similarity, accounting for lexical variation. |
| `Readability` (↓) | 1–3 | *1* = fluent, grammatically correct, easy to understand; *2* = includes invented words, errors, or scientific terms; *3* = as in 2, plus more scientific terms, making it harder to understand. |
| `Completeness` (↑) | 0–1 | *1* = full, concise answer. Relaxed: allows one incomplete sentence; Strict: at least one complete sentence. |
| `Correctness` (↑) | 0–1 | *1* = medically correct and in French. Relaxed: general meaning clear; Strict: exact meaning clear and complete. |

Table 4.6: Manual Evaluation Scoring criteria for readability, completeness, and correctness.

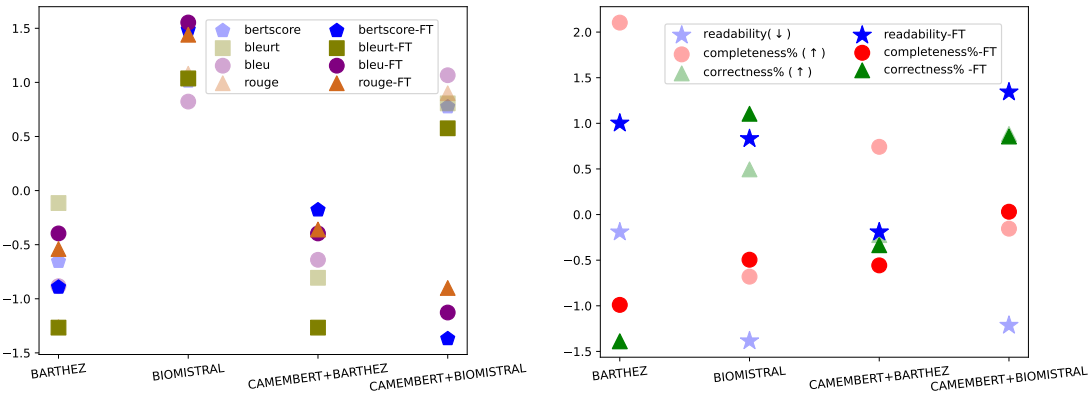
**Experimental Setup.** The experimental setup is build on baselines case study with BioMistral (BioMistral-7B-SLERP-GPTQ4[10]; Labrak et al. (2024)) and BARThez (Eddine et al., 2020) models in Section 3.2.3. Firstly, we test language models' ability to generate medical paraphrases in an zero-shot setting. Then, we compare these generated paraphrase with the knowledge base integrated settings, the RefoMed-KB corpus via our proposed pipeline pRAGe. We also test the two models in two settings: non-fine-tuned (NonFT) and fine-tuned (FT) on the RefoMed paraphrase dataset. For all configurations, we consider two different lengths for the generated text: 25 and 50 tokens. For the inference setting, we used a simple prompt in French (Figure 4.5) for the Base model and a RAG adapted prompt in French for our pRAGe pipeline (Figure 4.9). We decided to use the prompts in French, as initial test experiments with English prompts generated French-English text.

## Results & Discussion

We present the summarized evaluation in Figure 4.7 its corresponding details in Table 4.8-4.9. The complete evaluation report is in Table B.5. The automatic evaluation shows that BioMistral responses are more semantically and lexically related to gold paraphrase compared to BARThez. Further, we observe that both models benefit from finetuning. However, we notice that BioMistral pRAGe setups obtained lower scores with finetuning. On the contrary, BARThez pRAGe setups overall stay unaffected by fine tuning. This observation can be attributed to the fact that in pRAGe setups,

[10]This is the GPTQ quantized version of BioMistral.

models are conflated by the external knowledge base[11] whereas in the case of model only setup, the models are free to generate anything and therefore, can be prone to hallucinations.




(a) Automatic Evaluation.

(b) Manual Evaluation.

Figure 4.7: Evaluation for PRAGE setup.

The human annotation of 50 examples in 24 different configurations (1200 samples) shows that the fine-tuned version of BioMistral in inference alone and integrated in the pRAGe pipeline is the best model for short refomed-paper/answers (90% strict correctness). Base BioMistral is the best model for longer answers in inference alone setting (94% strict correctness).

The best model from all the configuration, BioMistral, non fine-tuned, generated an English word in a French sentence, as seen in example [1]. After finetuning, the model generates a correct answer in French [2][12]. In the context of the pRAGe pipeline, BioMistral gives as well a full French answer [3].

1. non fine-tuned - *Asthme: maladie où les airways se ferment et se contractent, faisant du bruitlors de l'inspiration et de la respiration* (Asthma: a disease in which the airways close and contract, making noise during inspiration and breathing)

2. fine-tuned - *maladies respiratoires chroniques et maladies rares respiratoires* (chronic respiratory diseases and rare respiratory diseases)

3. pRAGe (CamemBERT) - *Asthme: maladie qui fait ressentir des difficultés à respirer, souvent accompagnée de toux et de sifflements.* (Asthma: a disease that makes breathing difficult, often accompanied by coughing and wheezing.)

[11] It is interesting to note, that BioMistral and similar models when under such situation are posed with the challenge of memorization versus parametrization i.e. should they generate the response that completely or partially overrides the external knowledge or should be only consider external knowledge.

[12] The query contained a list of terms (*asthme , mucoviscidose , ventilation mécanique* - asthma , cystic fibrosis , mechanical ventilation), thus explaining the plural form of the generated text.

With respect to the constraint of 25 and 50 new tokens, the manual analysis shows that the best model is BioMistral accuracy in correct answers in a relaxed and strict setting (96% ; 94%), while the second best is its fine-tuned version (94% ; 90%). However, the model is not as good in strict correctness in a 25 token setting (68%). Our fine-tuned version of BioMistral is better (90%, up to 22% increase in performance) on strict correctness and conciseness in a 25 tokens setting. In the pRAGe pipeline, CamemBERT BioMistral, both non fine-tuned and fintuned, gave better answers in terms of strict correctness for the 25 token setting (82%; 81%).

For *readability*, in the short answer setting (*Token=25*), the *readability* is better with Base BioMistral (1.08, lower is better). However, even if the readability is good, the answers are incomplete (10% *completeness-strict*). Our fine-tuned version of Base BioMistral improves the completeness of the answer (16% `completeness-strict`), while the best pRAGe model, CamemBERT BioMistral fine-tuned, increases it even further up to 33%. Thus, we see a +23% improvement in performance (in *completeness-strict*) with the fine-tuned model in pRAGe.

One aspect that explains the decrease in *readability* in the fine-tuned models is the higher use of medical terms in the generated answers, as the fine-tuning step with the RefoMed dataset focuses on medical terms. As the aim of the medical text simplification task was to give short and concise paraphrases to a user query, we see that there are advantages of the fine-tuned model: it generates sub-sentential paraphrases, thus shorter and complete units of meaning (*Token=25*). Moreover, the fine-tuned model also generates simplified text generations, as observed in the following example where the medical term "osteophyte" is explained by using a subs-entential paraphrase in very simple language: "deposits of bone tissue that form on the edges of bones" (original in French *ostéophyte - des dépôts de tissu osseux qui se forment sur les bords des os*).

Additionally, we computed a Krippendorff's alpha score (Krippendorff, 2018) for two criteria of the human evaluation: *completeness* and *readability*. The annotation was conducted by 3 French linguists annotators: 1 linguist completed a full annotation and 2 other linguists contributed to the second annotation (one annotated Token=25 and the other Token=50 length paraphrases). We show the Krippendorff's alpha score and the percentage agreement in Table 4.7. The inter-annotator agreement is highest for *completeness-strict*, for both lengths (98% agreement), showing syntactic analysis is an easy task for the annotators. However, regarding *readability*, it is more difficult for the two annotators to agree (78% to 68% agreement), meaning that the medical knowledge of the annotators can influence the readability level of annotations.

| | **Krippendorff's alpha(nominal)** | **% agreement** |
|---|---|---|
| token=25 | | |
| **w/o FINE TUNING** | | |
| *completeness*-STRICT | 0.879 | 98% |
| *completeness*-RELAX | 0.649 | 90% |
| *readability* | 1.555 | 78% |
| **w/ FINE TUNING** | | |
| *completeness*-STRICT | -0.076 | 84% |
| *completeness*-RELAX | -0.1 | 80% |
| *readability* | 0.132 | 68% |
| token=50 | | |
| **w/o FINE TUNING** | | |
| *completeness*-STRICT | 0.66 | 98% |
| *completeness*-RELAX | 0.003 | 64% |
| *readability* | 0.105 | 70% |
| **w/ FINE TUNING** | | |
| *completeness*-STRICT | 0.105 | 70% |
| *completeness*-RELAX | 1.151 | 64% |
| *readability* | -0.529 | 2% |

Table 4.7: Inter-annotator agreement analysis

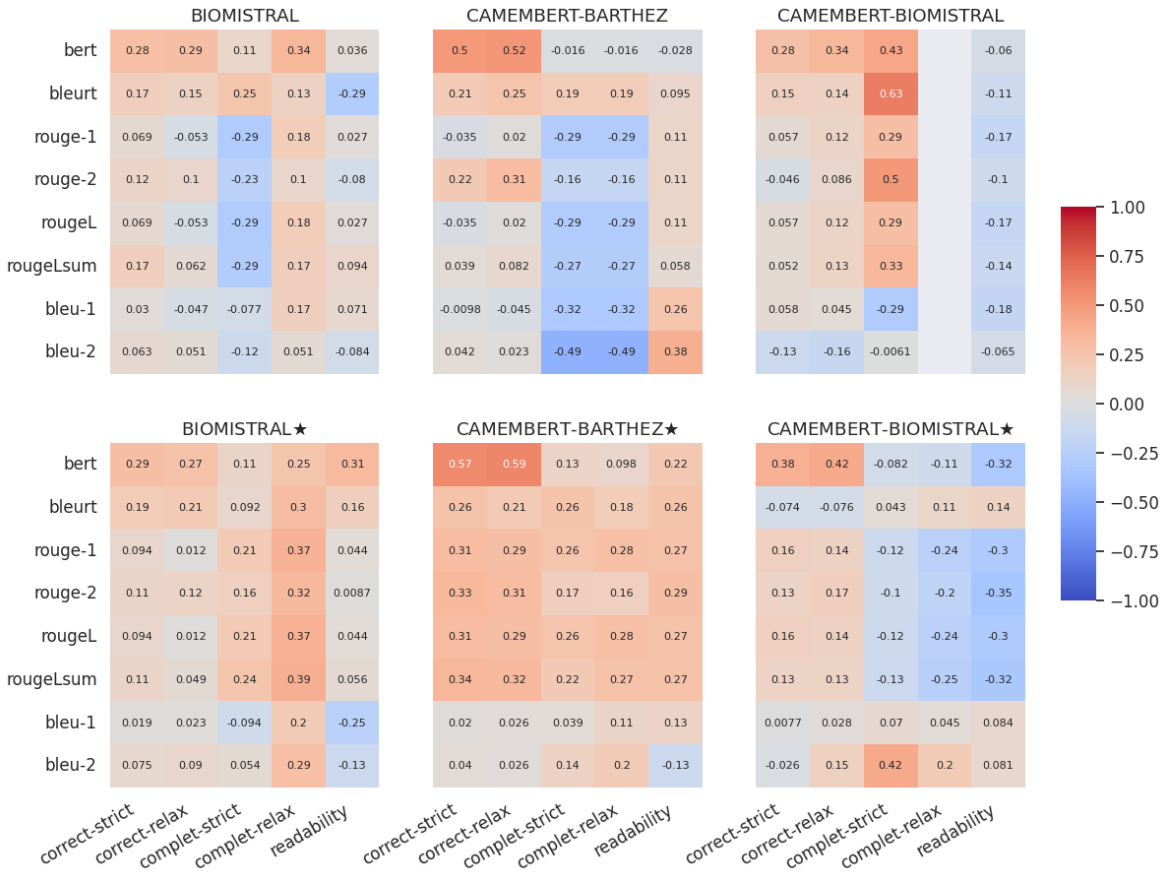



Figure 4.8: Correlation Heatmap between Automatic evaluation metrics (y-axis) and Manual evaluation metrics (x-axis). The ★ symbol denotes configurations with finetuned SLM.

Figure 4.8 combines automatic evaluation (Table 4.8) and manual evaluation (Table 4.9) with a linear correlation heatmap for best model and pRAGe configurations. The top row corresponds to configuration without finetuning (w/o FT) and the bottom row corresponds to configurations with finetuned (w/ FT) model backbone. ROUGE metrics positively corrected with *correctness* for the generation with finetuning which is intuitive as finetuning allows the pRAGe setups to generate in a more similar style as the gold reference (RefoMed). BERTScore and BLEURT scores remain unaffected with finetuning as the semantic similarity of generation can saturate unless lexical similarity increases. Finally, we notice the *readability* aspect of models is differently affected from pRAGe setup as it is not as trivial as lexical closeness.

## Summary

To conclude, in this work we presented pRAGe, a pipeline for *retrieval*, *generation*, and *evaluation* of medical paraphrases, designed to integrate external knowledge bases into language models. We applied this pipeline to the same models used previously in

Section 3.2.3, namely BARTHEZ and BioMistral, for the task of medical paraphrase generation.

Our proposed approach shows improvement for *scientific grounding* of text generation models in the medical domain, as evidenced by notable improvements in clinical correctness for model such as BARTHez with no prior knowledge (See table 4.9 green highlights). This aspect is particularly important given the high-risk nature of the medical domain. These results support the utility of integrating external knowledge bases to enhance the factual and domain-specific grounding of language models in healthcare applications.

However, there remains significant room for improvement in overall model performance. This suggests that, while multi-source modeling is beneficial, the specialized characteristics of the medical domain demand more than the addition of external resources—they require deeper *factual* and *semantic* understanding of the domain.

**Limitations** One key limitation of this study is that it includes only two models, which is insufficient to make general claims about other paraphrase generation models. Additionally, the study does not cover a broader range of large language models (LLMs), such as Vigogne (Huang, 2023) or Claire [13], in future. Another limitation is the lack of evaluation using more recent metrics designed for generative tasks, such as the *LM-as-a-Judge* (Gu et al., 2024) framework.

[13]https://huggingface.co/OpenLLM-France/Claire-7B-0.1

Vous êtes un expert en médecine. Utilisez les informations suivantes pour répondre à la question de l'utilisateur par une paraphrase, une explication ou une courte définition. Si vous ne connaissez pas la réponse, dites simplement que vous ne savez pas, n'essayez pas d'inventer une réponse.
**Contexte**: {context}
**Question**: {question}
Ne renvoyez que la réponse utile. La réponse doit être claire, concise et facile à comprendre pour le grand public.
**Réponse utile** :

You are a medical expert. Use the following information to answer the user's question with a paraphrase, an explanation, or a short definition.
If you do not know the answer, simply say that you do not know; do not attempt to make up a response.
**Context**: {context}
**Question**: {question}
Return only the useful answer. The response should be clear, concise, and easy to understand for the general public.
**Answer**:

Figure 4.9: Our prompt template for pRAGe setup (top = original in French and bottom = translation in English).

| | Model Setup | | Tokens=25 | | | | Tokens=50 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | bert | bleurt | bleu-1 | rouge-1 | bert | bleurt | bleu-1 | rouge-1 |
| | | | w/o FINE TUNING | | | | | | | |
| SLM | BARTHEZ | | $0.63_{0.03}$ | $0.10_{0.10}$ | $0.04_{0.06}$ | $0.07_{0.08}$ | $0.63_{0.03}$ | $0.10_{0.10}$ | $0.04_{0.06}$ | $0.07_{0.08}$ |
| | BIOMISTRAL | | $\mathbf{0.70}_{0.06}$ | $\mathbf{0.15}_{0.15}$ | $\mathbf{0.11}_{0.12}$ | $\mathbf{0.20}_{0.16}$ | $\mathbf{0.68}_{0.06}$ | $\mathbf{0.16}_{0.15}$ | $\mathbf{0.08}_{0.08}$ | $\mathbf{0.18}_{0.13}$ |
| pRAGe | camemBERT | BARTHEZ | $0.65_{0.05}$ | $0.07_{0.09}$ | $0.05_{0.07}$ | $0.12_{0.10}$ | $0.65_{0.05}$ | $0.11_{0.11}$ | $0.05_{0.06}$ | $0.12_{0.10}$ |
| | DrBERT | BARTHEZ | $0.64_{0.03}$ | $0.02_{0.06}$ | $0.04_{0.06}$ | $0.10_{0.09}$ | $0.65_{0.04}$ | $0.05_{0.07}$ | $0.05_{0.06}$ | $0.11_{0.09}$ |
| | camemBERT | BIOMISTRAL | $\mathbf{0.69}_{0.06}$ | $\mathbf{0.14}_{0.15}$ | $\mathbf{0.12}_{0.14}$ | $\mathbf{0.19}_{0.17}$ | $\mathbf{0.68}_{0.06}$ | $\mathbf{0.17}_{0.15}$ | $\mathbf{0.08}_{0.09}$ | $\mathbf{0.18}_{0.14}$ |
| | DrBERT | BIOMISTRAL | $\mathbf{0.69}_{0.06}$ | $\mathbf{0.14}_{0.15}$ | $0.11_{0.12}$ | $0.18_{0.17}$ | $\mathbf{0.68}_{0.06}$ | $\mathbf{0.17}_{0.15}$ | $\mathbf{0.08}_{0.08}$ | $0.17_{0.13}$ |
| | | | w/ FINE TUNING | | | | | | | |
| SLM★ | BARTHEZ | | $0.62_{0.02}$ | $0.05_{0.08}$ | $0.06_{0.07}$ | $0.11_{0.08}$ | $0.63_{0.03}$ | $0.09_{0.09}$ | $0.07_{0.07}$ | $0.12_{0.08}$ |
| | BIOMISTRAL | | $\mathbf{0.72}_{0.07}$ | $\mathbf{0.15}_{0.17}$ | $\mathbf{0.14}_{0.13}$ | $\mathbf{0.22}_{0.17}$ | $\mathbf{0.69}_{0.07}$ | $\mathbf{0.16}_{0.16}$ | $\mathbf{0.10}_{0.10}$ | $\mathbf{0.18}_{0.13}$ |
| pRAGe★ | camemBERT | BARTHEZ | $\mathbf{0.65}_{0.05}$ | $0.05_{0.09}$ | $\mathbf{0.06}_{0.07}$ | $0.12_{0.10}$ | $\mathbf{0.64}_{0.05}$ | $0.10_{0.10}$ | $\mathbf{0.06}_{0.07}$ | $\mathbf{0.12}_{0.10}$ |
| | DrBERT | BARTHEZ | $0.64_{0.03}$ | $0.01_{0.04}$ | $\mathbf{0.06}_{0.07}$ | $\mathbf{0.13}_{0.10}$ | $\mathbf{0.64}_{0.04}$ | $0.05_{0.07}$ | $0.05_{0.06}$ | $\mathbf{0.12}_{0.09}$ |
| | camemBERT | BIOMISTRAL | $0.60_{0.04}$ | $\mathbf{0.13}_{0.11}$ | $0.03_{0.03}$ | $0.09_{0.06}$ | $0.60_{0.05}$ | $\mathbf{0.16}_{0.11}$ | $0.03_{0.03}$ | $0.09_{0.06}$ |
| | DrBERT | BIOMISTRAL | $0.59_{0.04}$ | $0.12_{0.15}$ | $0.03_{0.03}$ | $0.08_{0.06}$ | $0.60_{0.04}$ | $0.14_{0.15}$ | $0.03_{0.02}$ | $0.08_{0.06}$ |

Table 4.8: Automatic Evaluation Metric Comparison of BaseSLMs with pRAGe setups on test set. Top scores for each model setups are shown in **bold**.

| | | | w/o FINE TUNING | | | | | w/ FINE TUNING | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *readability* (↓) | *completeness*%(↑) | | *correctness*%(↑) | | *readability* (↓) | *completeness*%(↑) | | *correctness*%(↑) | |
| | | | STRICT | RELAX | STRICT | RELAX | | STRICT | RELAX | STRICT | RELAX |
| **SLMs** | BARTHEZ | 1.22 | **100** | **100** | 0 | 0 | 1.36 | 0 | 0 | 0 | 0 |
| | | 1.20 | **100** | **100** | 0 | 0 | 1.42 | 0 | 0 | 0 | 0 |
| | BIOMISTRAL | **1.08** | 10 | 20 | 68 | **96** | **1.34** | 16 | 20 | **90** | **94** |
| | | 1.10 | 18 | 96 | **94** | **96** | 1.5 | **24** | **42** | **90** | **94** |
| **pRAGe** | camemBERT BARTHEZ | 1.22 | 56 | 64 | 42 | 46 | 1.22 | 14 | 14 | 38 | 42 |
| | | 1.26 | **96** | 96 | 46 | 50 | 1.34 | 70 | 76 | 48 | 50 |
| | DrBERT BARTHEZ | **1** | 18 | 68 | 0 | 0 | **1.04** | 60 | 60 | 0 | 0 |
| | | 1.46 | 94 | 94 | 0 | 0 | 1.08 | **90** | **92** | 0 | 0 |
| | camemBERT BIOMISTRAL | 1.10 | 27 | 33 | 82 | 88 | 1.40 | 33 | 48 | 81 | 90 |
| | | 1.06 | 37 | **100** | **88** | **90** | 1.56 | 10 | 33 | **90** | **92** |
| | DrBERT BIOMISTRAL | 1.04 | 14 | 24 | 46 | 84 | 1.20 | 34 | 38 | 74 | 88 |
| | | 1.08 | 32 | 98 | **88** | 88 | 1.50 | 14 | 32 | 72 | 88 |

Table 4.9: Manual evaluation comparison of BaseSLMs with pRAGe models

### 4.2.3 Case Study XI: External KG for Document Classification

This chapter is based on work previously published in our article: *Sinha, A., Bigeard, S., Clausel, M., & Constant, M. (2023). What shall we read: the article or the citations?-A case study on scientific language understanding. In Actes de CORIA-TALN 2023. Actes de l'atelier "Analyse et Recherche de Textes Scientifiques" (ARTS)@TALN 2023, pages 80–85, Paris, France. ATALA.* Parts of the text, figures, and results are adapted from this publication.

Representing data as a *graph* gained significant attention in natural language processing due to its ability to capture complex relationships and structured information that is often not explicitly available in raw text (Wu et al., 2023). Graph-based representations enable the modeling of intricate dependencies between entities, concepts, and their relational structures, providing a rich framework for encoding semantic information beyond traditional sequential or bag-of-words approaches. This paradigm shift led to the successful employment of knowledge graphs for a variety of NLP tasks, including recommendation, question answering, and classification (Mondal et al., 2021; Jin et al., 2019b), demonstrating their potential to encode rich semantic and relational information that enhances model performance and interpretability. Similarly, ontology-integrated models have shown that incorporating domain-specific hierarchical knowledge can improve both interpretability and task performance (Sinha et al., 2022).

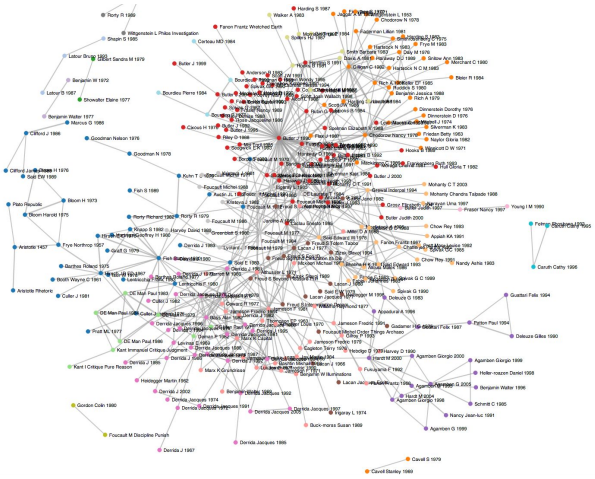


Figure 4.10: Illustration of citation Network, an example of knowledge graph.

Building on these insights, in this case study we explore the use of graph representation learning techniques (Perozzi et al., 2014; Hamilton et al., 2017; Kipf and Welling, 2017) to combine structural information from knowledge graphs with textual features. By leveraging the connectivity and semantic relations in graphs, models can potentially learn more robust representations that capture both local and global contextual dependencies, which are particularly relevant for complex tasks such as document-level understanding and information extraction.

**Research Question 1.2c:** *Can external knowledge graph improve LM's generalization for document understanding?*

## Methodology

**Task Description.** In this study, we revisit the scientific document classification task that was introduced in §3.4.1. This task involves a document used as input, and the language model has to classify it into one of the classes provided in the dataset. It is to be noted that here we focus on the utilization of an external knowledge graph, for eg. citation graph corresponding to the dataset, to complement the standalone baseline models introduced during the single-sourced model approach to study document classification.

**Dataset** We revisit the *PubMed-Diabetes* (Namata et al., 2012) which was introduced in the section §3.4.1. The dataset contains 19717 articles belonging to 3 classes of diabetes-mellitus, *Experimental*, *Type-1*, and *Type-2*. The dataset also provides citation information (eg. paper A $\xrightarrow{\text{cites}}$ paper B) and original PubMed-IDs. Using the graph terminology, PubMed-ID(s) denotes node(s) and the citation relation denotes edge(s) and therefore, the dataset can be also viewed as a citation graph with 19716 nodes[14] and 44338 edges.

**Models.** For this study, we use DeepWalk (DW) (Perozzi et al., 2014) and GraphSage (GS) (Hamilton et al., 2017) to generate graph embeddings corresponding to each article in the dataset using the citation information. Both models learn low-dimensional vector representations of nodes in a graph, yet they differ in how they leverage structural and neighborhood information.

(a) **DeepWalk** is an unsupervised algorithm that learns latent representations of nodes by simulating random walks over the graph. Each random walk is treated analogously to a sentence in a language model, where nodes correspond to words and co-occurrence patterns capture local structural proximity. The Skip-Gram model (Mikolov et al., 2013a) is then employed to learn embeddings such that nodes sharing similar neighborhoods are mapped to nearby points in the embedding space. This approach effectively captures community structure and homophily within the graph while remaining computationally efficient.

(b) **GraphSAGE** (Graph **S**Ample and aggre**G**at**E**) extends beyond random-walk-based methods by enabling inductive representation learning. Instead of relying on global random walks, it samples a fixed-size neighborhood for each node and learns to aggregate feature information from those neighbors using functions

[14]Note: One of the article-id (*pid*.17874530) was not anymore valid on PubMed, so we removed it from the nodes set; no article was connected with the removed article-id and so the edge list was unaffected.

such as mean, LSTM-based, or pooling aggregators. This inductive formulation allows GraphSAGE to generalize embeddings to previously unseen nodes, making it particularly suitable for large or evolving citation graphs. Moreover, by incorporating node-level attributes (e.g., textual or metadata features), GraphSAGE can combine topological and semantic information more effectively than purely structural methods.

And, in addition to these graph based features, we consider tf-idf feature that were manually generated TF-IDF vectors (tfidf-m) computed from specific text portions (e.g., title-only or abstract-only) for comparative analysis.

**Evaluation Metrics.** We evaluating the feature based models of document classification we used Macro F1 Scores across the different feature embedding models.

## Results & Discussion

Figure 4.11 represents the baselines results that were obtained in single sourced approach with standalone language models for scientific document models in Section 3.4.1.

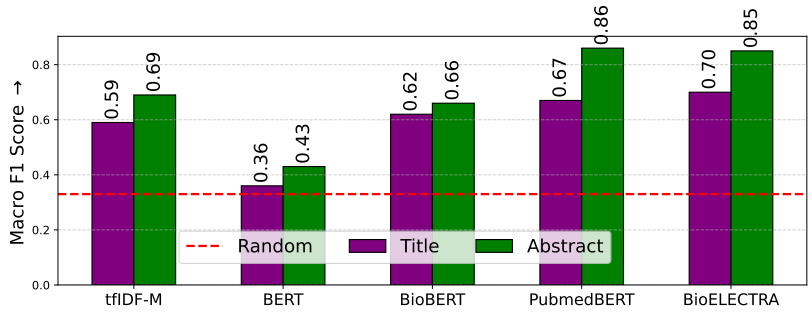



Figure 4.11: Performance comparison of language (embedding) models for Scientific Document Classification.

From fig. 4.12, we observe the performance of the *DeepWalk (DW)* model, noting that its informativeness is comparable to that of *tf-idf* features. We also experiment with injecting noise features as node inputs to the GraphSAGE (GS) model and observe a significant drop in performance compared to using static graph-based features (DW). This highlights the importance of high-quality initial node features for the GS learning algorithm.

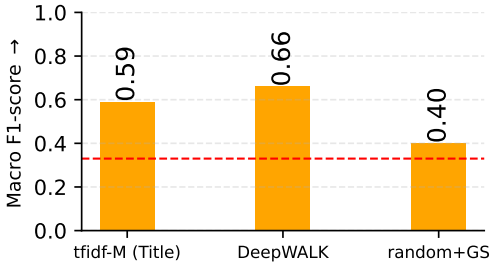



Figure 4.12: Comparison of graph-only features against text based features.

Interestingly, the *DW* features outperform *tf-idf-M* trained on **Title**-text (when compared to *tf-idf-M* alone) (See fig. 4.12), but

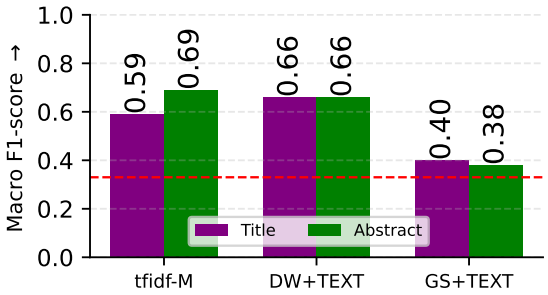



Figure 4.13: Performance Comparison between text-only feature verus text+graph features for scientific document classification.

not when trained on **Abstract**-text. This suggests that the relative informativeness of structural features depends on the quality and length of the underlying textual modality.

We begin by examining the concatenation of DW and tf-idf features (see fig. 4.13). Interestingly, tf-idf-M(**Abstract**) + DW performs comparably to tf-idf-M(**Title**) + DW, yet both combinations underperform relative to tf-idf-M(**Abstract**) alone. This suggests that the addition of graph-based features does not always enhance performance, particularly when the textual signal is already strong. Notably, DW alone outperforms tf-idf-M trained on **Title**-text (when compared to tf-idf-M alone), but not when trained on **Abstract**-text.

Next, we evaluate GS under two settings. In the first, where tf-idf-M features are used as node inputs, we observe a performance drop compared to the text-only setting. This may be due to the iterative message passing in GraphSAGE, which can degrade the original feature quality during node embedding generation. In the second setting, we use tf-idf-C features and observe improved performance over the tf-idf-M variant (cf. table 4.10). This improvement may be attributed to the curated nature of the keyword list and the sparsity of the tf-idf-M representation.

Overall, we find that combining text and graph features tends to outperform using either source individually in most cases. These results reinforce the importance of jointly leveraging both textual and structural information in downstream tasks.

## Summary

In this case study, we revisited the [SCI] document classification task introduced in Section §3.4.1, where we previously demonstrated that language models benefit from longer textual input for improved document understanding. Here, our focus was to investigate whether incorporating an external knowledge graph—specifically, a citation network—could enhance classification performance.

The results indicate that graph-based features, even when used standalone, perform comparably to traditional count-based text features for document classification.

Furthermore, we observed that a simple concatenation-based approach for combining textual and graph-based features leads to improved performance compared to using either feature set alone.

**Limitations** This study has several limitations. First, the experimental setup is constrained to traditional NLP features and classical graph representation learning algorithms. Recent advances, such as Graph Transformers (Yuan et al., 2025) and other Neural Graph Encoders (Wu et al., 2019), were not explored. Second, the feature fusion strategy was limited to straightforward concatenation; future work could investigate more sophisticated embedding fusion techniques to enable more effective multi-source integration.

| | Method | Title | | | | Abstract | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **P** | **R** | **F1** | **Acc** | **P** | **R** | **F1** | **Acc** |
| R | random | $0.36_{0.03}$ | $0.33_{0.03}$ | $0.33_{0.03}$ | $0.32_{0.03}$ | - | - | - | - |
| T | tfidf-c | $0.70_{0.02}$ | $0.68_{0.03}$ | $0.67_{0.03}$ | $0.68_{0.03}$ | - | - | - | - |
| | tfidf-m | $0.61_{0.03}$ | $0.60_{0.02}$ | $0.59_{0.03}$ | $0.60_{0.02}$ | $0.71_{0.02}$ | $0.70_{0.03}$ | $0.69_{0.03}$ | $0.70_{0.03}$ |
| G | DeepWALK (DW) | $0.67_{0.03}$ | $0.64_{0.03}$ | $0.65_{0.03}$ | $0.64_{0.03}$ | - | - | - | - |
| | random+GraphSAGE(GS) | $0.17_{0.10}$ | $0.36_{0.08}$ | $0.21_{0.08}$ | $0.36_{0.08}$ | - | - | - | - |
| T+G | tfidf-c+DW | $0.68_{0.02}$ | $0.65_{0.02}$ | $0.65_{0.02}$ | $0.65_{0.02}$ | - | - | - | - |
| | tfidf-m+DW | $0.68_{0.02}$ | $0.65_{0.02}$ | $0.66_{0.02}$ | $0.65_{0.02}$ | $0.69_{0.02}$ | $0.65_{0.02}$ | $0.66_{0.02}$ | $0.65_{0.02}$ |
| | tfidf-c+GS | $0.78_{0.03}$ | $0.76_{0.02}$ | $0.76_{0.02}$ | $0.76_{0.02}$ | - | - | - | - |
| | tfidf-m+GS | $0.42_{0.02}$ | $0.40_{0.04}$ | $0.40_{0.04}$ | $0.40_{0.04}$ | $0.41_{0.01}$ | $0.39_{0.02}$ | $0.38_{0.01}$ | $0.39_{0.02}$ |

Table 4.10: Comparison of Graph-based features against random and combination with text features for Scientific Document Classification.

## 4.3 Use of additional modality

Most deep learning models rely solely on source of data, typically sourced from social media posts, clinical narratives, or biomedical literature. While powerful, these single-sourced models are inherently limited by the information encoded in one source alone. In particular, medical decision-making is multimodal in nature, involving not only structured tabular data but also textual documentation, medical imaging, laboratory results, and physiological signals.

Incorporating additional modalities into medical models offers the potential to overcome several challenges faced by single-sourced medical systems. For example, radiology reports are written in relation to medical images, and understanding the report content may benefit from access to the corresponding scan. Similarly, integrating lab results or vital signs with clinical notes can improve context understanding for diagnostic reasoning or patient outcome prediction.

Modeling approaches combining multiple modalities enable medical models to learn richer, complementary representations of medical data. They can help disambiguate vague or incomplete measurement, improve robustness in data scarcity settings, and support more medically grounded predictions. Architectures for handling such data range from early fusion models that jointly embed multiple modalities, to late fusion systems that combine separate modality-specific predictions.

However, integrating additional modalities also presents significant challenges. These include data alignment and synchronization or inconsistent measurements, which may increase computational complexity, and the need for large-scale, multi-modal training datasets — many of which are not publicly available due to privacy concerns.

In the following case studies, we explore the extent to which additional modality can enhance medical NLP systems and examine their role in mitigating the weaknesses of single-sourced models. More concretely, we look into the following research question:

RQ 2.2

Can the use of additional modality help to mitigate the challenges faced by single sourced models?

### 4.3.1 Case study XII: Multimodal learning

Missingness is an ubiquitous phenomena that characterizes real time data. In particular, for time series type data, *missingness* denotes the unobserved data measurements. This can be because of various controllable and uncontrollable reasons. In the case of medical domain, we observe Missing Not At Random (MNAR) phenomena because often the decision of data measurements is taken by clinical experts. For eg., compared to a healthy patient, the clinicians would ask for more or frequent diagnosis for a ill patient. Such missingness is characterized by temporal pattern, i.e., irregularity (see fig. 4.14) that can be indicative of certain event that can be relevant for clinical decision making. In the literature, such missingness are termed as "informative missingness" (Rubin, 1976). However, such missingness phenomena are commonly mismanaged through imputation[15]. Furthermore, while structured data such as tabular physiological measurements have been the focus of various imputation techniques, the medical domain is not limited to such data types. Unstructured data, such as clinical notes, are also prevalent and are well-documented in the literature for their rich informational content. However, they pose significant challenges due to their heterogeneous and unstructured nature.

Furthermore, these types of missingness are not merely technical absences but are embedded in the language, style, and clinical reasoning found in narrative notes. This raises the question of whether additional, complementary sources of information such as other modalities could help models better interpret and learn from such underdocumented patterns.

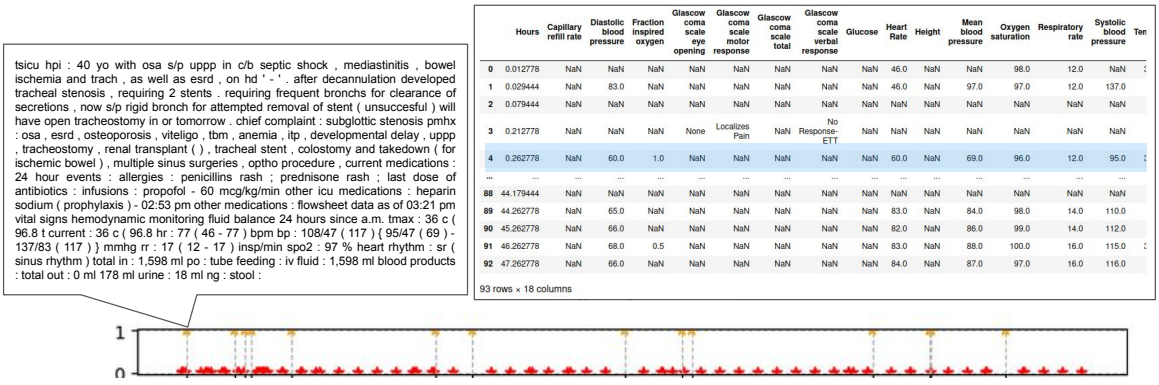



Figure 4.14: Illustration of co-existing modalities in clinical setting. * denotes physiological measurement variable and * denotes clinical notes.

With single sourced modeling in Section 3.5.1, we explored structured data, demonstrating how patterns of missingness can be used to deduce clinical decision-making processes without imputation. Building on that foundation, we now turn our attention to unstructured clinical narratives [CLIN], which are rich in contextual and temporal

[15]Imputation refers to the process of filling in missing data points using estimated values based on observed data. While convenient, imputation can introduce bias or distort the true data distribution, especially when the missingness mechanism is not well understood.

signals but inherently more complex due to their heterogeneity and lack of annotations. Our interest here is driven by the hypothesis that integrating an additional source of information such as a complementary modality may help improve the performance on clinical outcome prediction task which requires ability to capture temporal patterns. In order to do so, we aim to explore the following research question:

**Research Question 2.2a:** *Can we learn better from missingness in [CLIN] medical data by using additional modality?*

By exploring this, we begin to move from single-source approaches towards multi-sourced modeling approaches for more robust, context-aware modeling of medical data.

## Background

The integration of multimodal information in medical time series analysis has emerged as a critical research direction, with substantial contributions addressing the fusion of temporal physiological signals and static clinical records. Khadanga et al. (2019) demonstrated the efficacy of combining time series data with clinical records for in-hospital mortality prediction using the MIMIC-III dataset, employing a convolutional neural network (CNN) architecture built upon pre-trained word embeddings to capture clinical context. This foundational work established the potential for multimodal approaches in critical care applications. Building upon this, Zhang et al. (2023b) extended the multimodal paradigm by incorporating long short-term memory (LSTM) networks to model the temporal dynamics of clinical records alongside time-series data, utilizing a multitask learning framework that jointly optimized for prediction and representation learning through their Concatenation-Based Baseline (CONCAT) approach.

Recent advances have focused on sophisticated fusion mechanisms and architectural innovations for handling multimodal medical data. Deznabi et al. (2021) leveraged fine-tuned BioBERT representations for clinical text processing while maintaining LSTM-based temporal modeling, demonstrating improved performance through domain-specific language model adaptation. Yang and Wu (2021) introduced attention mechanisms to selectively focus on relevant multimodal information, while Yang et al. (2021) explored diverse fusion strategies including additive, concatenation, and multiplicative approaches, ultimately proposing a neural architecture search method to automatically discover optimal fusion configurations. Zhang et al. (2023b) addressed the critical challenge of irregular sampling in multimodal time series, developing specialized fusion modules that handle temporal misalignment and leverage cross-modal attention mechanisms to capture complex interactions between different data modalities.

The field has also witnessed significant contributions in handling missing data and improving model interpretability in multimodal settings. Niu et al. (2023) introduced fusion modules specifically designed to integrate multimodal features while preserving individual modal contextual information, enabling more robust mortality outcome prediction through enhanced feature correlation modeling between time series and clinical records. Advanced attention mechanisms have been explored by (Ilievski and Feng,

2017) and (Su et al., 2020), who proposed multimodal segmentation attention modules capable of processing feature blocks with varying spatial dimensions and sequence lengths, while maintaining compatibility with existing pre-trained unimodal networks. Contemporary work has expanded into specialized domains, with recent contributions from (Hayat et al., 2022; Yang and Wu, 2021; Raghu et al., 2022; Kline et al., 2022; Liu et al., 2022, 2023) providing comprehensive surveys on multimodal learning applications in medical domains, contrastive learning approaches, and domain-specific challenges. These works collectively highlight the evolution from simple concatenation-based fusion to sophisticated attention-driven architectures that can effectively leverage complementary information across temporal physiological signals, static clinical variables, and textual clinical narratives, establishing a robust foundation for next-generation multimodal medical AI systems.

# Methodology

**Problem Formulation.** We are interested in modeling multimodal time series which in particular belong to the category of irregularly sampled time series (ISTS). We denote a time series with $\mathbf{X}$ and its associated outcome with $Y$. On expanding, $\mathbf{X}$ comprises of multiple input channel $s$ (for eg. heart rate, systolic blood pressure, etc) and different modalities $m$ (such as tabular, text, images, etc).

For this case study, we build on SLAN model which was presented in Section 3.5.1. It included physiological tabular data $S$, we focus on Clinical Notes (CN) as an additional modality. We have a list of timestamps[16], input channels and clinical notes. SLAN (Agarwal et al., 2023b) presented an adaptive LSTM-based model for irregularly sampled physiological time series data that dynamically changes its architecture depending on the measured sensors ($m$) at any time point by utilizing a switch layer. For each input channel ($x_j^m$), it contains a separate LSTM block ($L^m$) which is activated only when the input channel is measured. Each LSTM block maintains a local summary ($h_j^m$) which is subjected to decay parameter using the input's last appearance ($\Delta_j^m$) to accommodate the irregularity importance.

$$(h_j^m, c_j^m) = L^m(x_j^m, h_{j-1}^m, c_{j-1}^m, \Delta_j^m) \; \forall m \in \mathbb{S}_j \tag{4.2}$$

Further, the information in any LSTM block is complemented by global summary via the aggregation cell at the time of its activation.

$$c_j = agg(\bigcup_{\forall m \in \mathbb{S}_j} \{c_j^m\}) \tag{4.3}$$

**Our Proposal : Multi-modal Switch LSTM Aggregation Network (mSLAN).** We integrate clinical note time series as another feature to the physiological time series data. Although, as compared to other temporal features, the dimension of clinical note

[16]We have aligned the timestamps of Time Series (TS) and CN by associating the Clinical note to the closest TS timestamp, and incase of multiple CNs for a timepoint we take the first one.

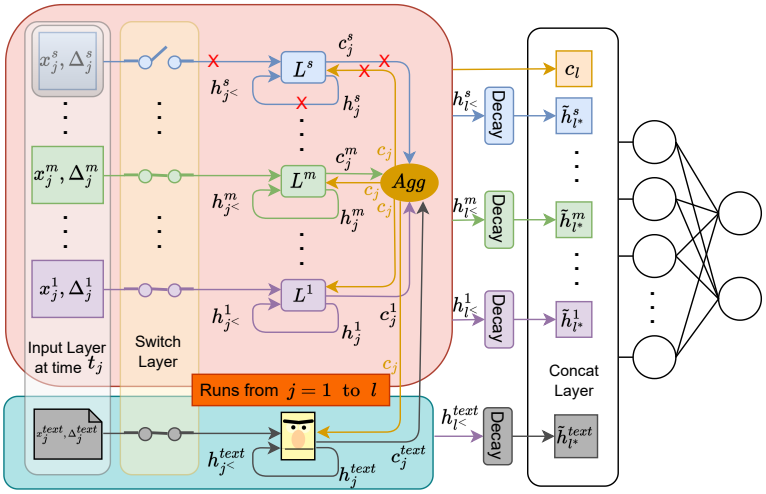



Figure 4.15: Snapshot of mSLAN Architecture.

encoding representation embedding would be 768 which will be required to project into smaller dimension before passing to the LSTM block. The interaction of information from various input channels happens in the aggregation layer to calculate $c_j$. We explore different type of aggregation method for inter modal feature interaction for example, vanilla concatenation as illustrated below to advanced fusion strategies (Sahu and Vechtomova, 2021).

$$c_j = agg(\bigcup_{\forall m \in \{\mathbb{S}_j, text\}} \{c_j^m\}) \tag{4.4}$$

Finally, the prediction is made by using a concat layer with local summaries from each LSTM and final global summary : A fully-connected network is employed to get a final prediction from $C$ as follows

$$\hat{y} = F(C) \tag{4.5}$$

**Dataset.** We utilize MIMIC (Johnson et al., 2016) dataset, same as the baseline case study in section 3.5.1. This is collection of reports of stays of patients in the critical care unit at a large tertiary care hospital. It has 21142 stays of unique patients instances with a median length of stay of 2.1 days. A total of 17 physiological measurements, like vital signs, medications, etc., are recorded for each patient. Following SeFT (Horn et al., 2020), we remove 32 instances which lead to the final number of instances decreasing to 21110.

**Task Description.** We utilize the in-hospital mortality prediction task for the current study, same as the baseline study in Section 3.5.1. The task involves binary classification that uses multivariate clinical time series data such as physiological tabular and clinical notes to a medical model as an input and the output from the model is expected to be 0 or 1 based on mortality prediction.

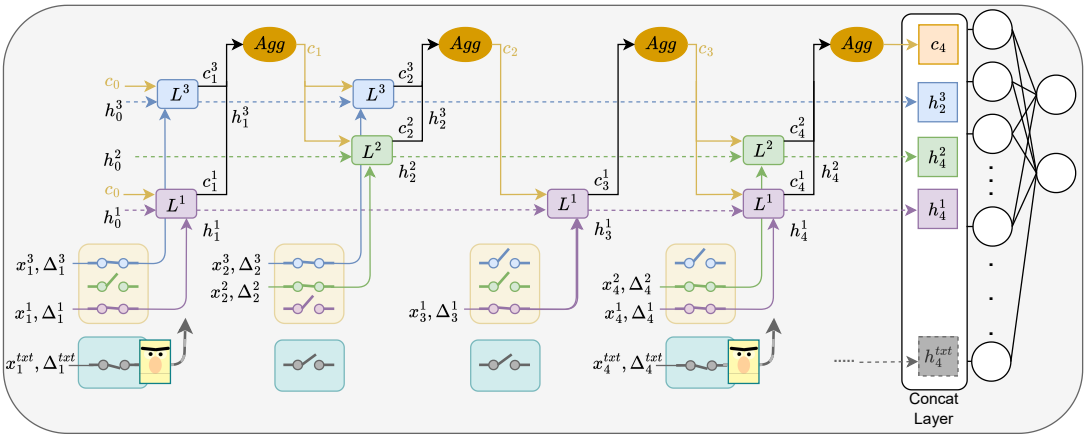



Figure 4.16: Unrolled illustration of mSLAN architecture

**Baselines.** We consider SLAN (referred to as Time Series SLAN (tsSLAN) this case study) introduced in the previous chapter (section 3.5.1) and Text-SLAN (txtSLAN)[17], for comparing the impact of different modalities wrt. mSLAN. txtSLAN operates same as SLAN network with multivariate. Further, we also utilize Unified Temporal Discretization-based Embedding (UTDE) [18](Zhang et al., 2023b) and all of its baselines to position mSLAN with respect to other multi-modal time series models. These baselines are multi-modal baselines such as Imputation (Zhang et al., 2023b), IP-Net (Shukla and Marlin, 2018), mTAND, GRU-D (Che et al., 2016) SeFT(Horn et al., 2020) RAINDROP (Zhang et al., 2021) DGM2-O (Wu et al., 2021b) and MTGNN (Wu et al., 2020), and text only baselines such as Flat, HierTrans, T-LSTM, FT-LSTM, GRU-D, and mTANDtxt.

**Fusion Techniques.** We consider four fusion techniques to combine tabular and clinical note features following the rollout illustration Figure 4.16:

1. **MEAN Fusion**: We compute the mean of the class probabilities obtained from the tsSLAN and txtSLAN models to produce ensemble-based predictions. This strategy represents a simple yet effective late fusion approach.

2. **EarlyFusion1**: Clinical notes are first encoded using a BERT+Low-Rank Adaptation (LoRA) encoder, followed by a linear projection to obtain reduced-length embeddings (e.g., lengths 32, 64, etc.) for each note, aligned with the clinical time series. These embeddings are then passed to the aggregator in the SLAN architecture, and the rest of the model functions as in the vanilla tsSLAN setup i.e. there is no separate SLAN unit used for text modality.

[17]txtSLAN is essentially the same tsSLAN architecture when provided input reduced lengthed embedding for irregularly sampled clinical text time series.

[18]UTDE embeds each modality (time series and clinical notes) onto a shared discretized timeline by dynamically gating between interpolation and attention embeddings for the tabular modality, and applying time-attention (mTAND) to note embeddings; fused via interleaved self- and cross-attention across modalities.

3. **EarlyFusion2**: Both time series and clinical text data are processed through their respective SLAN networks. We treat the reduced embedding of each clinical note as a multivariate time-series feature. For example, if the full embedding is projected to a dimension of 32, it is treated as a 32-dimensional feature vector. This converts the clinical text time series into a structured multivariate format. The representations from both networks are finally merged using their respective global summary vector $C$, and the concatenated output is used for final classification.

4. **LateFusionE2E**: Separate tsSLAN and txtSLAN networks are trained independently for each modality. Each network produces a global summary vector and local feature-level summaries. These outputs are concatenated and passed to an *adaptive fusion module*, which is implemented as an encoder-decoder architecture. The intermediate representation referred to as the fused vector is treated as the joint representation of both modalities and is used for the downstream classification task.

**Evaluation Metrics.** Following the baseline case study, we again use the area under the receiver operating characteristic (AUROC) and the area under the precision-recall curve (AUPRC) as comparison metrics among models.

# Results & Discussion

Table 4.11 presents the results of four different fusion strategies used to combine modalities in the mSLAN architecture. Each fusion approach yields improvements over the individual modality baselines. Notably, the late fusion (end-to-end) strategy achieves the highest performance gain relative to the tsSLAN baseline, highlighting the effectiveness of jointly leveraging both modalities during training and inference.

| | **AUROC** | **AUPRC** |
|---|---|---|
| MEAN | $87.4_{0.1}$ | $55.4_{0.4}$ |
| EarlyFusion1 | 87.8 | 53.9 |
| EarlyFusion2 | 88.3 | 57.3 |
| LastFusionE2E | **89.1** | **57.1** |

Table 4.11: mSLAN Fusion Experiments

Table 4.12 presents an ablation study comparing different modalities using the *SLAN architecture, where * denotes a modality-specific or multimodal variant of the SLAN model. The results demonstrate that even a simple combination strategy such as taking the mean of the sample-wise output probabilities from

| | **AUROC** | **AUPRC** |
|---|---|---|
| tsSLAN | $85.6_{0.1}$ | $51.1_{0.4}$ |
| txtSLAN | $85.9_{0.0}$ | $52.3_{0.0}$ |
| MEAN | $87.4_{0.1}$ | $55.4_{0.4}$ |
| mSLAN | $89.1_{0.0}$ | $57.1_{0.0}$ |

Table 4.12: Comparison of modality combination with various *SLAN architecture.

the tsSLAN and txtSLAN models yields improvements over both unimodal models. Furthermore, we observe additional performance gains when mSLAN is trained end-to-end with both modalities jointly, highlighting the benefit of full multimodal integration.

| | **Imputation** | **IP-Net** | **mTAND** | **GRU-D** | **SeFT** | **RAINDROP** | **DGM2-O** | **MTGNN** | **UTDE** | **mSLAN** (Ours) |
|---|---|---|---|---|---|---|---|---|---|---|
| **AUPRC** | $44.36_{1.36}$ | $39.36_{1.10}$ | $47.54_{1.28}$ | $45.90_{0.40}$ | $23.89_{0.46}$ | $36.23_{0.37}$ | $37.79_{1.54}$ | $36.49_{2.10}$ | $49.64_{1.00}$ | $57.1_{0.0}$ |

Table 4.13: UTDE and its baseline compared against best mSLAN variant.

| Model | Flat | HierTrans | T-LSTM | FT-LSTM | GRU-D | mTANDtxt | txtSLAN (Ours) |
|---|---|---|---|---|---|---|---|
| AUPRC | $51.69_{0.79}$ | $52.98_{1.69}$ | $52.57_{3.25}$ | $54.39_{1.38}$ | $54.34_{0.75}$ | $56.05_{1.09}$ | 52.3 |

Table 4.14: Performance Comparison for clinical notes modality (AUPRC).

Table 4.13 presents a comparison between our proposed model, a best variant of mSLAN and various baselines from UTDE (Zhang et al., 2023b). Further Table 4.14 presents an ablation analysis of mSLAN comparing its performance against other models also using only the clinical note modality. The multi-modal results show that the MEAN mSLAN fusion variant outperforms all other multi-modal time series models considered however, txtSLAN obtains comparable results. However, it is important to note that a direct comparison with the UTDE baselines is not entirely fair due to differences in the datasets: outliers[19] were removed from the dataset used for the mSLAN experiments, resulting in a slightly different number of test samples (2482 for mSLAN versus 2488 for UTDE). Despite this, the results provide a useful indication of the potential standing of the mSLAN model.

## Summary

In this study, we revisit the temporal characteristics of medical data, previously introduced in Section §3.5.1, with a focus on examining the impact of incorporating an additional modality namely, clinical notes for processing data with missing values through imputation. Our proposed model, mSLAN, integrates clinical text as a complementary modality and demonstrates improved performance over the baseline SLAN model, which relies solely on physiological tabular data for modeling missingness. These findings further support the intuition that multi-source modeling of medical phenomena such as missingness can provide complementary information, leading to more robust and accurate representations.

**Limitations** The primary limitation of this study is the exclusion of more advanced time-series architectures, such as point process-based approaches, which may also be

[19]The description of removal of outliers is provided in the Appendix B.1.

well-suited for modeling missingness phenomena in medical data.

## 4.4 LLM as an Assistant

Large Language Models (LLMs) have demonstrated remarkable capabilities across a wide range of natural language tasks, including question answering, summarization, reasoning, and even few-shot adaptation to unseen problems (Brown et al., 2020; Chowdhery et al., 2023). In the biomedical and clinical domain, domain-adapted variants such as BioGPT (Luo et al., 2022), PubMedBERT (Gu et al., 2021b), and Med-PaLM (Singhal et al., 2023a,b) have shown strong performance on benchmark datasets and practical tasks. However, LLMs are not without limitations, particularly when used as standalone systems in high-stakes, data-sensitive domains such as healthcare. Key concerns include hallucination which is, generation of plausible but factually incorrect information (Ji et al., 2023; Zhang et al., 2023e) – as well as potential biases, lack of transparency, and limited ability to handle rare or complex clinical scenarios (Thirunavukarasu et al., 2023).

An emerging paradigm explores the use of LLMs not as autonomous agents, but as assistants to existing models or systems (Moor et al., 2023). In this context, LLMs can provide complementary reasoning, generate candidate outputs for downstream filtering, assist with ambiguous cases, or enhance robustness in zero- or few-shot settings (Labrak et al., 2023d; Alsentzer et al., 2023). This assistant-based usage frames the LLM as a support mechanism that can be called upon selectively, rather than a monolithic model that replaces all other components. For example, in named entity recognition pipelines, LLMs have been employed to validate or correct borderline predictions from task-specific models or in the case of clinical decision support systems, LLMs may suggest relevant literature passages or generate explanations based on structured patient data through retrieval-augmented generation. The modularity of this approach offers flexibility and the opportunity to combine the complementary strengths of multiple systems.

In the following case study, we investigate the potential of large language models to serve as supportive components in medical NLP pipelines and examine whether their integration can mitigate known weaknesses of single-sourced models.

RQ 2.3

Can the use of large language model (LLM) as an assistant help to mitigate challenges faced by Single-sourced models?

### 4.4.1 Case study XIII: Overcoming Writing Variation

Clinical notes serve as a cornerstone for evidence-based decision-making, supporting prognostic reasoning and the evaluation of treatment outcomes (Pham et al., 2017). However, unlike structured medical records, clinical notes are composed in free-text by diverse healthcare professionals whose writing styles, terminologies, and syntactic habits vary widely. This stylistic heterogeneity shaped by institutional norms, individual training, and time constraints introduces significant variability in narrative structure, from the use of shorthand and omitted subjects to irregular sentence construction and domain-specific phrasing.

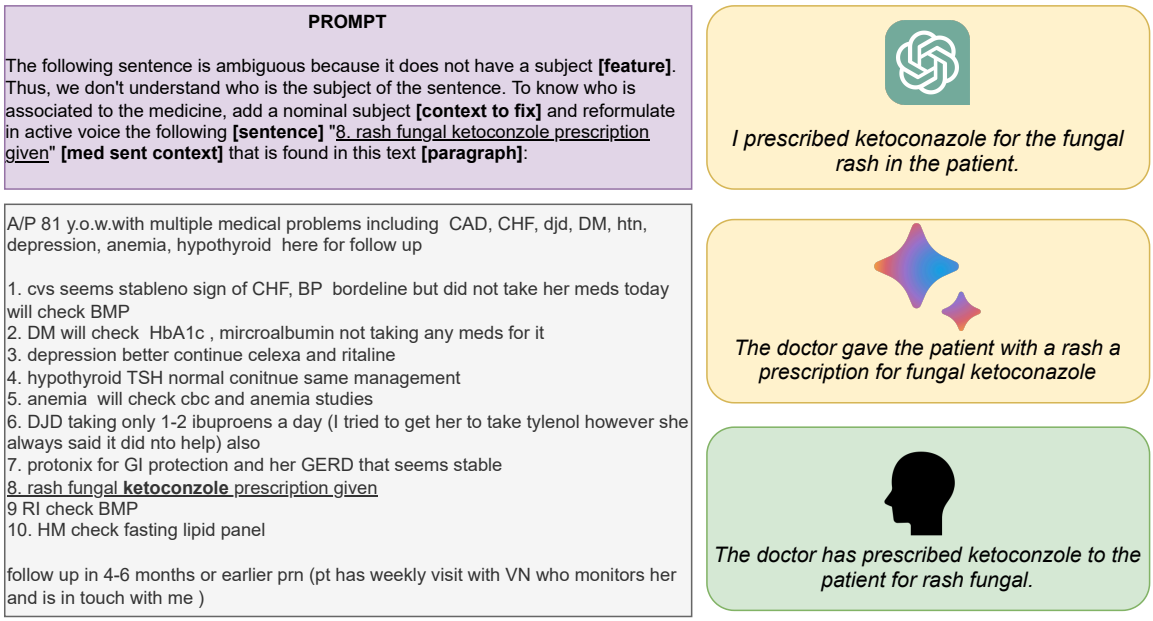



Figure 4.17: Illustration of LLM as an assistant. Bottom-left box is an example of a instruction prompt for a target medicine (eg. ketoconzole) in the underlined sentence in a clinical note. Right top first and second box denotes LLM generated explanation.

Such variation poses a substantial challenge for NLP models, which are prone to learning surface-level cues and author-specific patterns rather than generalizable representations. As shown in prior work (Shah et al., 2020a), this can result in poor performance across demographic or institutional subgroups and reduced model robustness. Standard contextualized language models, even those pre-trained on biomedical corpora, may struggle to reconcile such linguistic inconsistencies.

In this case study, we explore another multi-source modeling strategy, where a general-purpose large language model is used as an auxiliary source of information to enhance the robustness of clinical NLP systems. We frame our question as below :

**Research Question 2.2a:** *Can LLMs be used as assistant to overcome writing style variation in [CLIN] medical data?*

More concretely, we leverage the LLM's zero-shot reasoning and contextual under-

standing capabilities to act as an *helper* assisting a downstream classifier in interpreting clinical text for the task of medication change event classification. This setup exemplifies how LLMs can serve as a complementary source to mitigate the overcome writing style variation.

## Background

Style variation is a form of linguistic variation has been studied for describing and explaining variation in linguistic form across society (Nguyen et al., 2021). Stylistic Vagueness or variation is very well acknowledged in medical writing (Mitchell, 2007) and is advised to avoid as much as possible in medical communication. Although, the existence of any linguistic variation can be attributed to demographics of international doctors across the world, majority of whom have English as their second-language (Verma et al., 2016). Linguistic variation between user and the content of clinical notes already has studied for medical dialog systems (Llanos et al., 2016b). However, the effect of the style variation have not been studied thoroughly for language models for medical context understanding.

Clinical notes are narrative and longitudinal, capturing a patient's medical history, assessments, and care plans across time. These texts vary widely in structure and writing style depending on the provider, institution, and note type (Pivovarov and Elhadad, 2015). Events such as medication changes are often embedded in nuanced context such as temporal cues, attribution, and intent that must be accurately captured for tasks like medication reconciliation (Cadwallader et al., 2013). While clinical event extraction has received significant attention, the effect of writing style variation on the contextual understanding of language models remains underexplored (Xu et al., 2022).

The longitudinal and narrative characteristic of clinical notes can be attributed to past events leading to the current presentation, and outline plans for future management depending on how the patient responds to prescribed treatments. They describe actions taken by the patient, the patient's family, and other providers involved in the patient's care. Extraction of medication changes from clinical notes must be accompanied by the necessary clinical context (e.g., when the change was introduced, who initiated the change) to be useful for real-world applications such as medication reconciliation (Cadwallader et al., 2013).

We consider counterfactual explanation (Wachter et al., 2017) as a mean to verify the impact the different linguistic feature that contribute to writing style variation in clinical notes on medical model's capabilities for contextual medical event extraction task. There are various works proposed inline with this approach such as POLYJUICE (Wu et al., 2021a), MICE (Ross et al., 2021) which provide a pipeline for counterfactual generation and evaluation.

# Methodology

**Dataset.** We re-use CMED (Mahajan et al., 2022) dataset that was utilized earlier for clinical writing variation (See section 3.3.2). This dataset came from the Partners HealthCare Electronic Medical Records (EMR). EMR at Partners HealthCare comprises a platform shared by two large academic tertiary hospitals – Massachusetts General Hospital (MGH) and Brigham and Women's Hospital (BWH), USA. The clincal notes are associated to the patients who were admitted for the Type II Diabetes (T2D).

**Contextual Medical Event Extraction.** The contextual medical event extraction was introduced in n2c2 challenge[20] (2022). It involved two-fold tasks, firstly, it requires to identify medicine for which any *medication change* is being discussed in the clinical note and secondly, additional contextual information associated the medication change has to be extracted. Medication change, is defined as "any discussion about a medicine change" and is termed to as , *disposition*, for a given patient mentioned in the clinical note. Further, for any medication change that is identified as *disposition*, the model is expected also has extract contextual information regarding the `Action`, `Negation`, `Temporality`, `Certainty` and `Actor` that are associated to for each identified medication change (See Figure 4.18).

In this case study, we are interested to utilize LLM as an assistant to generate explanation for clinical narrative such that it can improve models with respect to contextual medical event extraction classification task.

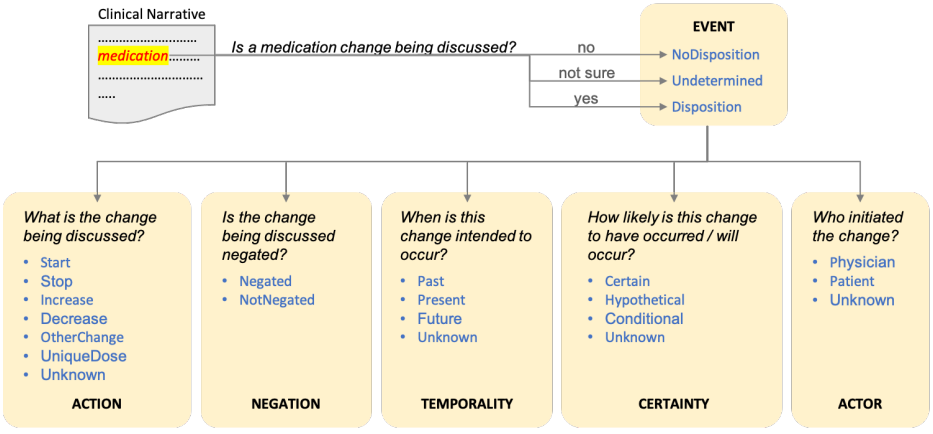



Figure 4.18: Detailed illustration of Contextual Medication Event Extraction (Mahajan et al., 2022).

**Linguistic characterization of Clinical Narrative.** We make use of BEA linguistic feature extractor tool (Lee and Lee, 2023) in order to extract linguistic features from

[20] https://n2c2.dbmi.hms.harvard.edu/

the segmented sentence of the clinical notes. The tool facilitates to extract for every sentence 220 linguistic feature. These linguistic features cover various linguistic characteristics on sentence-level such as surface, syntax, lexico-semantics, and discourse. We utilized SciSpacy[21] additionally instead of general domain modules for processing the segmented sentences.

**Models.** We consider four pre-trained language models as follows: BERT (Devlin et al., 2018), ModernBERT (Warner et al., 2024), ClinicalBERT (Huang et al., 2019) and PubmedBERT (Gu et al., 2021b). For each of these BERT-like models, we train classification model using the segmented sentences containing the target medicine as an input against each of the five context category labels for 20 epochs with early stopping. Details for reproducibility in Table 4.19.

**Algorithm 1** Counterfactual Evaluation

**Require:** A sentence $s$, target medicine $m$ from a dataset $D$, a linguistic feature $l \in \mathrm{L} = \{\varnothing, l_1, l_2, ..\}$ and a model $\mathcal{M}$.

- ▶ Train model $\mathcal{M}$ on $D(\equiv D(\varnothing))$.
- ▶ Collect predictions on $\hat{Y}_{\mathcal{M},D(\varnothing)} \equiv \hat{Y}_{\varnothing}$

**for** $l \in \mathrm{L} \setminus \{\varnothing\}$ **do**

- ▶ For every $s$, using $llm$ obtain $s_l$.
- ▶ $D(l) \leftarrow s_l$

**end for**

- ▶ Collect predictions $\hat{Y}_{\mathcal{M},D(l)} \equiv \hat{Y}_l$
- ▶ Compare distribution $\mathrm{Q}(\hat{Y}_{\varnothing})$ and $\mathrm{Q}(\hat{Y}_l)$.

**Algorithm.** We present the algorithm utilized in the experiment in algorithm 1. For any medical sentence $s$ containing a target medicine $m$ taken from a dataset $D$ we consider a set of linguistic features L that contains linguistic features $l_i$ and medical model trained for contextual event classification $\mathcal{M}$. Firstly, using the trained model $\mathcal{M}$ we obtain prediction (denoted by $\hat{Y}_{\mathcal{M},D(\varnothing)}$) on the original test data $D(\varnothing)$ where $\varnothing$ implies that no linguistic features are modified yet. Next, we utilize an $llm$ model and with each linguistic feature $l$ , we obtain an explanation (or modification) of the original sentence $s$ and denote it by $s_l$ repeating it for the entire dataset to obtain $D(l)$. After that, we obtain the prediction on the modified dataset $D(l)$ using the same model $\mathcal{M}$ to obtain the new predictions $\hat{Y}_{\mathcal{M},D(l)}$ or $\hat{Y}_l$. And finally, we compare the difference between $\hat{Y}_{\varnothing}$ and $\hat{Y}_l$.

**Experimental Setup.** Each clinical note is segmented at the sentence level, and only those sentences containing a medication change annotated with a *disposition* label are retained. We extract 220 linguistic features for selected disposition based sentences, and then use the models listed in table 4.17 to obtain predictions on the development

[21] en_core_sci_sm

and test sets. To test our hypothesis, we first define sample-level *correctness* by comparing ensemble predictions of all the four models with gold labels. We then with the obtained predictions, we apply the OLS method (Kuchibhotla et al., 2019) to examine correlations between linguistic features and correctness across the five context categories.

Based on obtained correlation report with OLS method and heuristics[22], we select a subset of features listed in Table 4.15 and use an LLM[23] that modifies validation and test sentences via prompting (cf. fig. 4.19) to alter linguistic features for counterfactual evaluation (cf. algorithm 1). Finally, the modified sentences are used for inference with the trained models to verify the changes in the correctness of predictions.

| Linguistic Feature ($lf$) | **Medical Sentence** | **Modifed Sentence** | **BEA**($s$) | **BEA**($s_l$) |
|---|---|---|---|---|
| #unique_nouns | –check chem 7 given recent restart BP meds | –check chemistry 7 given recent restart blood pressure medications | 3 | 5 |
| simple_verbs_variation | 5 x2 and lopressor 12. | 5 multiplied by 2 and lopressor 12. | 0 | 1 |
| #unique_adverbs | Pain was relieved by Morphine (4, 4, 2 mg). | Pain was completely relieved by Morphine (4, 4, 2 mg). | 0 | 1 |
| root_verbs_variation | 8. Neurontin 200 mg b.i.d. (she was told she could discontinue | Neurontin 200 mg b.i.d. (she was informed she could cease taking it). | 1.414 | 1.732 |
| #named_entities_person | We will have her start the Norvasc and Cozaar, have her return in two weeks for followup of blood pressure. | We will have Jane start the Norvasc and Cozaar, have Jane return in two weeks for followup of blood pressure. | 0 | 2 |
| simple_type_token_ratio | She was treated with IV lopressor 2. | She received IV lopressor twice. | 1.125 | 1.2 |
| avg_syllables_per_word | A1c off it 3/94 8. | A1c off it three over ninety-four eight. | 0.667 | 1.25 |

Table 4.15: Samples showcasing linguistic features modification via GPT-3.5-Turbo.

| | **BEA**($D$) | **BEA**($D(l)$) | **Action** | **Actor** | **Certainty** | **Negation** | **Temporality** |
|---|---|---|---|---|---|---|---|
| #unique_nouns | 2.9442 | 3.5535 | neg | pos | pos | pos* | pos |
| simple_verbs_variation | 0.8865 | 0.9463 | pos* | neg | neg | pos* | neg |
| #unique_adverbs | 0.6000 | 1.0419 | pos* | pos | pos | pos* | pos |
| root_verbs_variation | 1.2275 | 1.3415 | neg* | pos | pos | neg* | pos |
| #named_entities_person | 0.2744 | 0.6512 | - | pos | - | - | - |
| simple_type_token_ratio | 1.1047 | 1.1262 | neg* | pos* | pos | neg* | pos |
| avg_syllables_per_word | 1.4065 | 1.6680 | pos* | pos | neg* | pos* | neg |

Table 4.16: Mean shift for selected linguistic features and their impact on context categories. * denotes non-significant, – denotes no correlation.

# Results & Discussion

Table 4.15 presents examples of linguistic modifications generated with an LLM for each of the selected linguistic features ($lf$), while **BEA**(..) score indicates the the linguistic

[22]From among the entire list of significant linguistic features, we performed the linguistic feature selection by prompting GPT5 with the list of features and asked to provide a subset of features that are simple and easier understand , followed by final manual selection.

[23]We use GPT-3.5-Turbo for modifying linguistic features in the sentences.

feature extractor value for that row feature $l$ before and after the modification. Next, using the selected features, we measure and provide the overall feature shift shown in Table 4.16 as $D(l)$ compared to the original distribution $D$. Table 4.17 shows the ensemble results and using the modified dataset for $D(l)$ for each $l$ it shows the difference of results on the modified test set with respect to the linguistic feature $l$ with color coding across the table with expected outcome indicated by green, contradicting outcome indicated by red and neutral indicated by gray..

| Models | **Action** | **Actor** | **Certainty** | **Negation** | **Temporality** |
|---|---|---|---|---|---|
| ENSEMBLE | 52.5000 | 50.2000 | 50.2000 | 49.5000 | 48.5000 |
| #unique_nouns | -0.7 | +1.0 | -1.9 | ±0.0 | +1.5 |
| simple_verbs_variation | +2.1 | -6.3 | +5.2 | ±0.0 | +0.3 |
| #unique_adverbs | -1.0 | +0.9 | -1.9 | ±0.0 | -1.7 |
| root_verbs_variation | -0.7 | -3.0 | +0.6 | ±0.0 | -1.3 |
| #named_entities_person | -0.3 | -2.2 | ±0.0 | ±0.0 | +0.1 |
| simple_type_token_ratio | -0.2 | -6.8 | +3.8 | ±0.0 | +1.5 |
| avg_syllables_per_word | -2.7 | -4.1 | -1.2 | ±0.0 | +2.6 |

Table 4.17: Ensemble context classification results (Macro F1) under counterfactual feature shift on the test set. Green indicates expected outcomes, red indicates unexpected outcomes, and gray denotes excluded cases.

While the ensemble baseline performs consistently across tasks, counterfactual shifts introduce both gains and losses depending on the linguistic feature are modified (See Table 4.17). Complete results are in the Table 4.20. Concretely, we obtain overall ∼60% cases as we expected with Action and Negation being expected behavior throughout. *Negation* remains relatively stable across all feature shifts, however for *Actor*, *Certainty* and *Temporality*, we obtain mixed behavior. *Actor*, *Certainty* are the most affected showing significant performance change under *simple_verbs_variation* and *simple_type_token_ratio* respectively. Conversely, *Action* and *Negation* seems to experience least changes on when subjected to linguistic feature shift. These results suggest that linguistic variations can both degrade and enhance contextual understanding suggesting a more non trivial phenomena, with *Actor* being particularly vulnerable to syntactic and lexical changes.

Overall, we propose the use of LLM as an assistant to help single-sourced medical models overcome writing style variation. However, it is also important to note that the linguistic modification performed by LLM often contained hallucination (see table 4.18) which should be monitored to be mitigated.

| **Linguistic Feature** ($lf$) | **Raw Sentence** | **Hallucinated Modification** |
|---|---|---|
| #unique_nouns | 5 x2 and lopressor 12. | 2 unique nouns are in 5 x2 and lopressor 12. |
| simple_verbs_variation | Dr. Prater decreased atenolol bid. | Dr. Prater reduced atenolol twice a day. |
| #unique_adverbs | STARTED ON 0. | STARTED ON 1. |
| #named_entities_person | Patient on Keppra in past-needed now? Will ask neurology. | Patient on Keppra in past-needed now? Will ask Dr. Smith. |
| avg_syllables_per_word | NO ATTACKS ON COLCHICINE. | NO AS-SAULTS ON COL-CHI-CINE. |

Table 4.18: Examples of LLM generated explanation of clinical context containing hallucinations.

**Summary** In this case study, we explored the use of large language models (LLMs) as an external source of information for multi-sourced modeling. We addressed the problem of writing style variation in clinical notes in the context of contextual event classification and evaluated the use of an LLM as an assistant to help medical models overcome style-related variability. Our experiments revealed an interesting finding: the LLM-generated explanations led to changes in the model's behavior, influenced by the corresponding linguistic feature characteristics. This approach could also contribute to mitigating the issue of limited clinical data.

**Limitations** Some limitations of this case study are as follows: First, it considers only a single large language model (LLM) to explore the use of LLMs as assistants. As a result, extending this work to include other LLMs is left for future research. Additionally, the feature selection process used for LLM explanation generation could be improved by incorporating more linguistically informed methods.

| | |
|---|---|
| Batch Size | 4 |
| Learning Rate | 1e-5 |
| Loss Function | Focal Loss |
| EPOCHS | 20 |
| Early stopping | True |
| Patience | 3 |
| MAX LEN | 128 |

Table 4.19: Hyperparameter Setup.

| Models | **Action** | **Actor** | **Certainty** | **Negation** | **Temporality** |
|---|---|---|---|---|---|
| BERT | 51.8355 | 44.7025 | 52.8036 | 49.5399 | 48.3012 |
| MdoernBERT | 43.5511 | 35.8259 | 42.5367 | 63.9017 | 53.5846 |
| BioClinicalBERT | 49.1310 | 50.3754 | 48.7667 | 49.5399 | 50.1213 |
| PubmedBERT | 71.4808 | 49.9539 | 53.6503 | 49.5399 | 39.3746 |
| ENSEMBLE | 52.5000 | 50.2000 | 50.2000 | 49.5000 | 48.5000 |
| #unique_nouns | 51.8000 | 51.2000 | 48.3000 | 49.5000 | 50.0000 |
| simple_verbs_variation | 54.6000 | 43.9000 | 55.4000 | 49.5000 | 48.8000 |
| #unique_adverbs | 51.5000 | 51.1000 | 49.5000 | 49.5000 | 46.8000 |
| root_verbs_variation | 51.8000 | 47.2000 | 50.8000 | 49.5000 | 47.2000 |
| #named_entities_person | 52.2000 | 48.0000 | 50.2000 | 49.5000 | 48.6000 |
| simple_type_token_ratio | 52.3000 | 434.0000 | 54.0000 | 49.5000 | 50.0000 |
| avg_syllables_per_word | 49.8000 | 46.1000 | 49.0000 | 49.5000 | 51.1000 |

Table 4.20: Ensemble context classification results (Macro F1) under counterfactual feature shift on the test set (Complete Results).

You are an assistant supporting a linguistic study.
Your task is to rewrite the given sentence by applying the following linguistic feature: {`feature`},
Return only the modified sentence, without additional text or explanation.
Sentence:

Figure 4.19: Prompt Template for LLM as an Assistant.

## 4.5 Conclusion

This chapter examined how medical language models can benefit from multi-sourced signals, extending the analysis beyond the single-sourced paradigms discussed in the previous chapter. The goal for this chapter was to evaluate whether enriching models with additional source of information such as external knowledge, complementary modalities, or large language model assistance could mitigate the data-centric limitations observed earlier.

The first set of case studies explored the integration of external knowledge bases and knowledge graphs. These experiments demonstrated clear improvements in entity extraction, text simplification, and document classification by grounding the single-sourced model with curated knowledge base and structured ontologies. However, they also revealed new challenges related to knowledge coverage, temporal validity, and the non-trivial balancing of latent features.

The second approach introduced additional modality, showing that multimodal learning can capture richer clinical and contextual cues to capture more information. Tabular–textual integration complements each other to improved robustness and representation quality in missingness setting, yet highlighted practical barriers such as data alignment, and model interpretability.

The final case study proposed a new paradigm positioning large language models as assistant, an alternate source of information rather than standalone predictors. When guided appropriately, LLMs were able to provide stylistic variation normalization and suggest adaptation to diverse textual conventions, leading to a promising hybrid paradigm where general-domain models support specialized medical models. Nevertheless, these gains depend heavily on careful prompt engineering with hallucination mitigation, and alignment with clinical reasoning.

Taken together, the findings suggest that multi-sourced medical models offer a compelling path toward greater robustness and contextual understanding by addressing the data related issues. By leveraging heterogeneous data and external structure, they begin to bridge the gap between computational prediction and expert-level knowledge. Yet, the complexity of integrating multiple knowledge sources hints us towards model-related issues due to its own risks from inconsistency and knowledge drift to loss of interpretability.

These observations set the stage for the next chapter, which turns from improving medical model through knowledge integration toward a reflective yet evaluation standpoint. Specifically, we ask whether the enriched representations observed here translate into medical-expert aligned behavior that is, whether medical model based reasoning can approximate the behavior of medical experts.

# Part III

# Revisit: Where Words Meet Medicine

# Chapter 5

# Medical Models vs. Medical Experts



In this chapter, we offer a reflective and systematic examination of the key challenges involved in deploying medical models in real-world settings. The discussion is structured around two complementary perspectives. First, we address model-centric challenges, focusing on how design choices must account for uncertainty, and how the pursuit of improved performance often involves increasingly complex model architectures. Second, we explore data-centric issues, highlighting how limitations in data representation can lead to a disconnect between conceptual understanding and subjective interpretations ultimately contributing to gaps in how clinical tasks are differently perceived by models and human experts.

Taken together, these perspectives reveal the complex interplay of technical and epistemological barriers that continue to shape medical modeling research. The chapter concludes with a discussion on the broader implications and future directions in the field.

## 5.1 The Proficiency Gap

The main hurdle for the integration of medical (AI) models into clinical workflows lies in the persistent gap between modeling systems and trained medical experts. Despite considerable progress in predictive performance and the ability of deep learning–based models to surpass traditional machine learning approaches, these systems remain far from replicating the nuanced decision-making skills of clinicians in real-world practice (Rajpurkar et al., 2022). This gap is particularly evident in tasks requiring uncertainty awareness, causal inference, and the incorporation of complex reasoning that extends beyond what is encoded in training data (Albano et al., 2025; Bedi et al., 2025).

Recent literature has underscored that while AI models may achieve impressive accuracy on benchmark datasets, their performance often deteriorates when applied to heterogeneous or unseen clinical scenarios (Hurd et al., 2013; Shah et al., 2020b). Medical experts, by contrast, are able to generalize across contexts, incorporate uncertainty into decision-making, and adapt their reasoning dynamically based on evolving patient conditions (Sendak et al., 2020b). Such differences highlight a structural asymmetry: models are optimized for statistical correlation, whereas clinicians integrate experience, domain knowledge, and patient-specific factors into diagnostic and therapeutic judgments (Castro et al., 2020; Rajkomar et al., 2018a; Sendak et al., 2020a).

Moreover, concerns regarding interpretability and trust remain central to this gap. While models can provide probabilistic outputs, clinicians must translate such outputs into actionable decisions that consider ethical, contextual, and interpersonal dimensions of care (Holzinger et al., 2017; Doshi-Velez and Kim, 2017; Tonekaboni et al., 2019). The mismatch between the statistical nature of model reasoning and the holistic reasoning of clinicians creates challenges not only for accuracy but also for accountability and adoption in practice.

In this chapter, we will reflect on the capability of existing medical models in order to understand this proficiency gap, situating it within both methodological and epistemological debates in medical AI. Specifically, we aim to discuss the areas which are of important concern for practical deployment of medical models such as handling uncertainty, interpretability of reasoning, alignment with human experts and difficulty perception. This leads us to the central research question of this chapter:

RQ 3

Can we explain the proficiency gap between Medical Models and Medical Experts?

## 5.2 Model Centric Issues

Understanding the gap between the performance of medical models and the proficiency of medical experts remains a central concern in the development of trustworthy clinical AI. Despite advances in domain-specific language models and large-scale pretraining, medical NLP systems often fall short of human-level reasoning, especially in high-stakes clinical settings. This gap is a key reason why such models continue to face skepticism regarding their deployment in real-world healthcare environments.

In this section, we adopt a model-centric perspective to explore how internal model attributes and behaviors can help explain and potentially reduce this performance gap. Rather than focusing solely on external tasks or datasets, we examine intrinsic model qualities that contribute to (or hinder) reliable performance. This includes factors such as the model's ability to quantify its own uncertainty, and its capacity for interpretability and explanation.

These characteristics are essential for building confidence in model predictions, particularly in clinical use cases where wrong or unexplainable outputs can have serious consequences. Uncertainty awareness allows models to signal when they may not be confident in a prediction, enabling safer human-in-the-loop workflows. Interpretability, on the other hand, supports transparency and error analysis by offering insights into how or why a model arrives at a particular output.

In the following two case studies, we investigate how these properties uncertainty estimation and interpretability can be used to better understand and potentially bridge the proficiency gap between medical models and clinical experts.

RQ 3.1

What are the issues related to the modeling practices contributing to the gap?

### 5.2.1 Uncertainty awareness of Medical Models

This chapter is based on work previously published in our article: *Sinha, A., Mickus, T., Clausel, M., Constant, M., & Coubez, X. (2024, August). Domain-specific or Uncertainty-aware models: Does it really make a difference for biomedical text classification?. In Proceedings of the 23rd Workshop on Biomedical Natural Language Processing (pp. 202-211).* Parts of the text, figures, and results are adapted from this publication.

In the general domain, deep learning models are typically trained in a data-driven manner with the primary objective of maximizing prediction accuracy (Goodfellow et al., 2016). However, focusing *solely* on predictive accuracy is insufficient in mission-critical applications. For instance, in a binary classification setting, both a softmax probability of 51% and 95% may yield the same predicted class. Yet, the latter reflects higher confidence and lower entropy, indicating a more reliable prediction, whereas the former, despite leading to the same decision, is considerably less trustworthy. Reliance on maximum predicted probability alone can thus introduce several well-documented pitfalls, ranging from limitations in domain generalization (Daume III and Marcu, 2006) to the propagation of social and systemic biases (McCoy et al., 2019; Schnabel et al., 2016).

Furthermore, these limitations often compound, leading to a severe deterioration of model performance in out-of-domain (Out-of-Distribution (OOD)) test scenarios (Hurd et al., 2013; Shah et al., 2020b). This challenge has motivated significant engineering efforts toward developing models tailored to specific domains, ranging from legal applications (Paul et al., 2023) to mission-critical settings (Lee et al., 2020; Singhal et al., 2023b). In the latter case, particularly within the medical domain, despite considerable enthusiasm among practitioners for data-driven and AI-based models, there remains substantial skepticism about fully trusting their predictions without uncertainty awareness. This key factor that plays an important role and adds to the hurdle of deployment of medical (AI) models is owed to the fact of accounting uncertainty in order to abstain from a decision, whenever required. However, there is no systematic guidelines to follow for designing medical models in order to overcome this gap.

The research area of domain adaptation suggests continuing training of pre-trained models (initially trained on a general domain) on target domains such as the biomedical domain, which has been shown to be quite helpful for producing domain-specific models for biomedical applications (Poerner et al., 2020). However, domain-specific models, while useful, are rarely considered a definitive solution, because increased performance does not always translate into reliability: deep learning models can be overconfident yet incorrect (Yuksekgonul et al., 2023). In the biomedical domain, it is crucial to have reliable models — in particular, insofar as accounting for uncertainty in prediction is concerned. For example, in the case of a risk scoring model used to rank patients for liver transplant, uncertainty-awareness becomes critical.

For a deep learning model to be uncertainty-aware, it must not only provide predictions but also produce a quantifiable analogue of "*I am not sure*," reflecting the confidence associated with those predictions. In practice, this requires generating a dis-

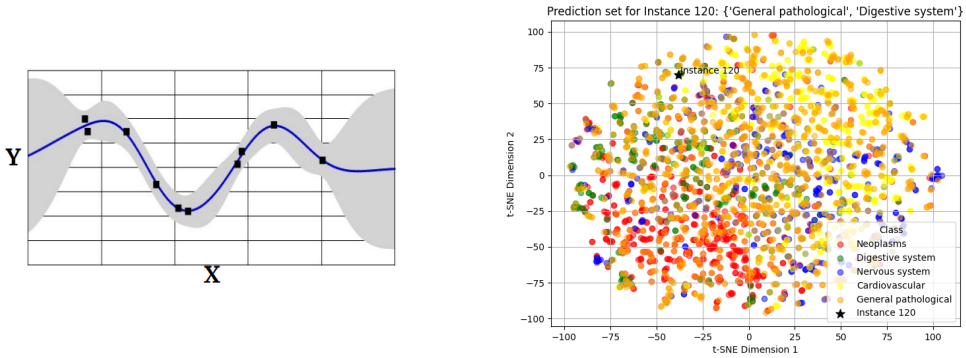



Figure 5.1: Uncertainty awareness in medical models

tribution over predictions rather than a single point estimate. As illustrated in fig. 5.1, for a regression model this may take the form of confidence intervals or prediction intervals, while for a classification model it may involve suggesting multiple plausible classes instead of committing to a single label. For instance, in fig. 5.1 (right), Instance 120, denoted by ★, is assigned two candidate classes General Pathological, Digestive System. This research area, known as *Uncertainty Quantification*, advocates the development of models that generate not only predictions but also an accompanying measure of reliability.

Uncertainty quantification has attracted increasing attention from the NLP community (Xiao and Wang, 2019a; Xiao et al., 2022; Hu et al., 2023), particularly in mission-critical domains such as medicine (Hwang et al., 2023; Barandas et al., 2024). In parallel, biomedical NLP has shifted from domain adaptation strategies (Wiese et al., 2017) toward domain-specific pretraining, with models ranging from BioBERT (Lee et al., 2020) to large-scale efforts such as MedPaLM (Singhal et al., 2023b). While Xiao et al. (2022) provide a detailed study of uncertainty paradigms in *general-domain* PLMs, such analyses overlook the distinct challenges that arise when domain specificity and uncertainty-awareness intersect. Previous work on uncertainty in biomedical AI has largely focused on structured data such as imaging or EHRs (Begoli et al., 2019; Abdar et al., 2021), leaving biomedical textual data comparatively underexplored. This gap is significant: medical language models often achieve high predictive accuracy, but without mechanisms for uncertainty-awareness they risk overconfident errors, potentially leading to misallocation of critical medical resources (Steyerberg et al., 2010). Taken together, these two lines of research, uncertainty quantification and domain-specific pretraining have progressed in parallel but rarely in dialogue. Whether their benefits are additive, overlapping, or mutually reinforcing remains unclear. This naturally leads to the central question of our study:

**Research Question 3.1a:** *Are the benefits of uncertainty-awareness and domain-specificity in LMs orthogonal? Should medical practitioners prioritize domain-specificity or uncertainty-awareness?*

In practice, we reflect on how model-specificity and uncertainty-awareness articulate with one another. Figure 5.2 illustrates the experimental setup we use for our study.

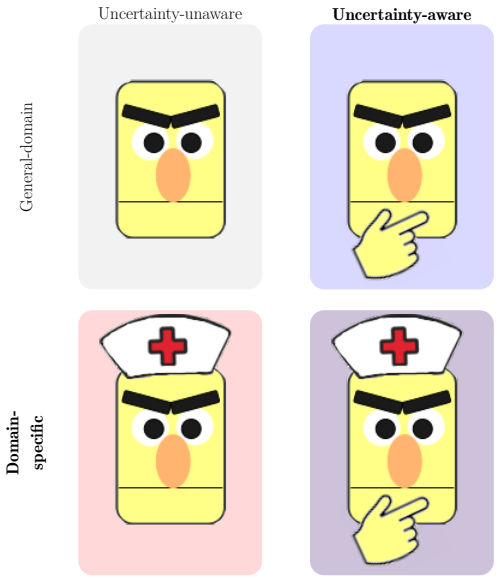



Figure 5.2: Uncertainty awareness experimental setup.

We study the performances of frequentist and bayesian general and domain-specific models on biomedical text classification tasks across a wide array of metrics, ranging from performance metrics to uncertainty quantification metrics.

## Methodology

**Formulation.** When training a biomedical model $f(X, Y)$ under a supervised setup with dataset pairs $(X, Y)$, where $Y$ is a discrete variable such that $y \in Y \subseteq C$, the model learns to map inputs to one of $C$ possible classes. At test time, inference yields a raw logit vector $f(x_{\text{test}}) \in \mathbb{R}^C$, which can be transformed into a probability distribution using a smoothing function (e.g., softmax), resulting in $\hat{p}(x_{\text{test}}) \in [0, 1]^C$. From this probability vector, we can obtain different quantifiers of uncertainty either at the sample level or aggregated across the dataset.

**Dataset.** We employ six biomedical datasets to investigate our research question. Three of these MedABS (Schopf et al., 2022a), MedNLI (Romanov and Shivade, 2018), and SMOKING (Uzuner et al., 2008) are in English, while the remaining three MORFITT (Labrak et al., 2023c), PxSLU (Kocabiyikoglu et al., 2022), and MedMCQA (Labrak et al., 2023b) are in French. An overview of the datasets is provided in Table 5.1 along with class imbalance ratio (CIR [1]; Yu et al. (2022)), with further details available in Appendix B.1. For MedABS, SMOKING, PxSLU, and MedMCQA, we use the raw input without additional preprocessing. In MedMCQA, we formulate the task as predicting the number of correct responses (ranging from 1–5) for each multiple-choice question. For MedNLI, we concatenate the premise and hypothesis with a [SEP] token and cast the task as multi-class classification. Finally, for MORFITT, which is

[1] $\text{CIR} = \frac{\text{Number of majority class samples}}{\text{Number of minority class samples}}$

originally multi-label, we use only the first label of each sample to convert it into a multi-class task.

| Dataset | Task Description | Splits | | | Statistics | | | |
|---|---|---|---|---|---|---|---|---|
| | | train | val | test | #Class | CIR | avglen | maxlen |
| MedABS | Predict the patient condition described, given a medical abstract | 8662 | 2888 | 2888 | 5 | 3.1445 | 180.59 | 597 |
| MedNLI | Predict the inference type, given a hypothesis and a premise | 11232 | 1395 | 1422 | 3 | 1 | 23.83 | 151 |
| SMOKING | Predict the patient smoking status, given a medical discharge record | 398 | 100 | 104 | 5 | 23.75 | 654.30 | 2788 |
| PxSLU | Predict the drug prescription intent, given a user speech transcription | 1386 | 198 | 397 | 4 | 98.1538 | 11.40 | 48 |
| MedMCQA | Predict the number of answers, given a medical multi-choice question | 2171 | 312 | 622 | 5 | 21.1176 | 12.90 | 92 |
| MORFITT | Predict the speciality, given a scientific article abstract | 1514 | 1022 | 1088 | 12 | 15.3529 | 226.33 | 1425 |

Table 5.1: Datasets description. CIR refers to class imbalance ratio.

**Models.** As illustrated in Figure 5.2, we compare the combination of two characteristics of medical models, domain specificity ($\mathcal{D}$) and uncertainty awareness ($\mathcal{U}$). For example, a domain specific uncertainty unaware model group can be denoted as ($+\mathcal{D}$ $-\mathcal{U}$). We derive classifiers from language specific PLMs: for English datasets, we use BERT (Devlin et al., 2019) and BioBERT (Lee et al., 2020); for French, we use CamemBERT (Martin et al., 2019) and CamemBERT-BIO (Touchent et al., 2023). We compare two types of models, frequentist Deep Neural Network (DNN) models and Bayesian Neural Network (BNN) models. The DNN model comprises of a PLM-based encoder, a Dropout unit along with 1-layer classifier. The BNN models are likewise based on a PLM encoder, along with a Bayesian module applied over the classification layer.

We experimented with Monte Carlo Dropout (MCD) models (Gal and Ghahramani, 2016), DropConnect (Mobiny et al., 2021), and Variational Inference (VI) (Blundell et al., 2015) models. We focus on the DropConnect architecture which comprises a PLM encoder along a DropConnect dense classification layer. This approach infuses stochasticity into a deterministic model by randomly zeroing out classifier weights with a probability $1-p$, where $p$ is the softmax probability. This allows us to sample multiple outputs for a given input, thus enabling to aggregate the predictions and to produce estimates of uncertainty.

**Evaluation Metrics.** We evaluate classifiers on two aspects: task performance and uncertainty awareness. For text classification, we report Macro-F1 and accuracy. For uncertainty quantification[2] we report Brier Score (BS) (Brier, 1950), Expected Calibration Error (ECE) (Naeini et al., 2015),Static Calibration Error (SCE) (Nixon et al.,

[2]We use the term *uncertainty-awareness* to broadly refer to a model's ability to represent and manage uncertainty in its predictions. This includes, but is not limited to, *calibration*—the extent to which predicted probabilities reflect true outcome likelihoods. For instance, a well-calibrated model that assigns 70% confidence to a class should be correct about 70% of the time. For this chapter, as the focus is on probabilistic outputs or confidence scores, the terms *uncertainty-awareness* and *calibration* may be used interchangeably.

2019), Negative Log-Likelihood (NLL), coverage (Cov%) and entropy (H). In what follows, $N$ denotes the number of samples in test set, $C$ denotes the number of classes. Lower score for Brier score, ECE, SCE, NLL and Entropy metrics; and higher score for coverage, are indicative of better uncertainty aware model.

1. **Brier score** (Brier, 1950) proposed BS which computes the mean square difference between the true classes and the predicted probabilities. A lower score is better, and a perfect *calibrated* model would have a Brier score of 0.

$$\text{BS} = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}(y_c^{(i)} - \hat{y}_c^{(i)})^2$$

2. **Expected Calibration Error** (Naeini et al., 2015) provides weighted average of the difference between accuracy and confidence across $B$ bins.

$$\text{ECE} = \sum_{b=1}^{B}\frac{n_b}{N}|\text{acc}(b) - \text{conf}(b)|$$

   where $\text{acc}(b)$ and $\text{conf}(b)$ are the average accuracy and confidence of predictions in bin $b$, respectively. We set $B = 15$ in our experiments.

3. **Static Calibration Error** Nixon et al. (2019) proposed an extension of ECE to multi-class problems to overcome its limitation of dependence of the number of bins.

$$\text{SCE} = \sum_{c=1}^{C}\sum_{b=1}^{B}\frac{n_b}{NC}|\text{acc}(b) - \text{conf}(b)|$$

   We set $B = 15$ in our experiments.

4. **Negative Log Likelihood** serves as the primary approach for optimizing neural networks in classification tasks. Interestingly, this loss function can also double as an effective metric for assessing uncertainty.

$$\text{NLL} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

5. **Coverage Percentage** quantifies the proportion of instances in which the true class label is contained within the predicted set. It is especially relevant for set-valued or top-$k$ predictions:

$$\text{Coverage} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\left[y_i \in \hat{Y}^{(i)}\right] \tag{5.1}$$

   where $\hat{Y}^{(i)}$ is the prediction set for sample $i$, and $\mathbb{1}[\cdot]$ is the indicator function.

6. **Shannon Entropy** (Shannon, 1948) quantifies the expected uncertainty inherent in the possible outcomes of a discrete random variable.

$$H = -\sum_{i=1}^{N} p_i \log(p_i)$$

**Reproducibility.** In all cases, we fine-tune the PLM backbone for all the downstream task with a maximum sequence length of 512 and a batch size of 50 sentences. We perform a hyper-parameter grid search for `epochs`= {3,4,5, ..., 15} and `lr`= {1e-7, 5e-6, 1e-6, 5e-5, 1e-5, 5e-4, 1e-4}. We replicate training with 3 seeds for each hyperparameter configuration, select the optimal configuration for validation F1, and replicate training with 7 more seeds for these optimal configurations, so as to obtain 10 models per dataset, PLM and architecture. We also select the main BNN model of the study by selecting the system yielding the highest average rank across all six datasets, as displayed in Appendix Figure 5.5. We train all models with binary cross-entropy loss and `Adam` optimizer with $\epsilon = 10^{-8}$ and $\beta = (0.9, 0.999)$. For all BNN models, we obtain 3 sets of predictions after training the models to calculate the mean class probabilities. Corresponding optimal hyper-parameters are listed in table B.9.

## Results & Discussion

When it comes to classification metrics (see fig. 5.3), models with domain-specific pre-training ($+\mathcal{D}$) generally outperform those without it ($-\mathcal{D}$). Since performance varies significantly across datasets, we first apply $z$-normalization[3] per dataset to remove overall trends and simplify comparison.

Overall, $+\mathcal{D}+\mathcal{U}$ models perform well, though they are sometimes outperformed particularly by $+\mathcal{D}-\mathcal{U}$ models on classification tasks. This pattern is especially clear in MedABS and MedNLI, where all $+\mathcal{D}$ models surpass their $-\mathcal{D}$ counterparts in both F1 and accuracy. In PxSLU, however, the $+\mathcal{D}-\mathcal{U}$ model shows notably lower accuracy, breaking the trend. For the two French datasets and SMOKING, the gap between domain-specific and general models is less pronounced. Full results are provided in Table B.10.

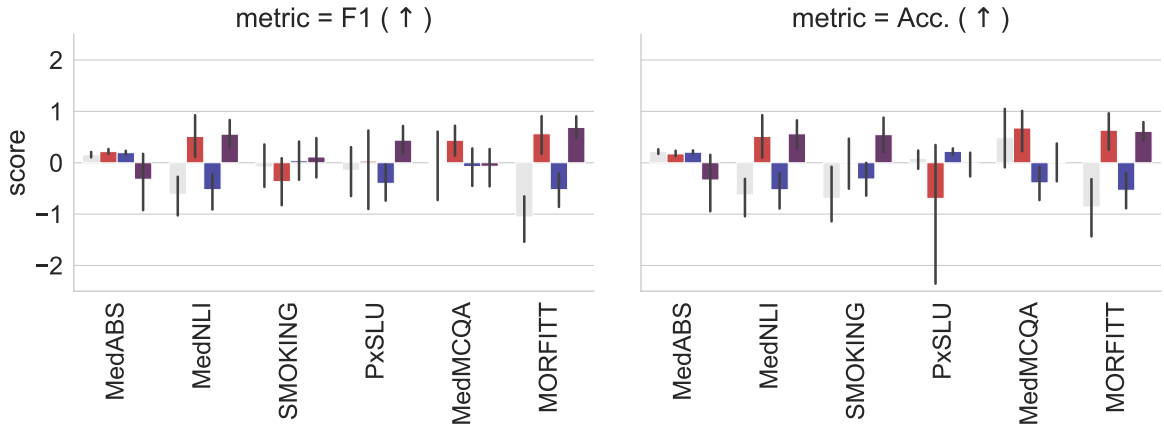
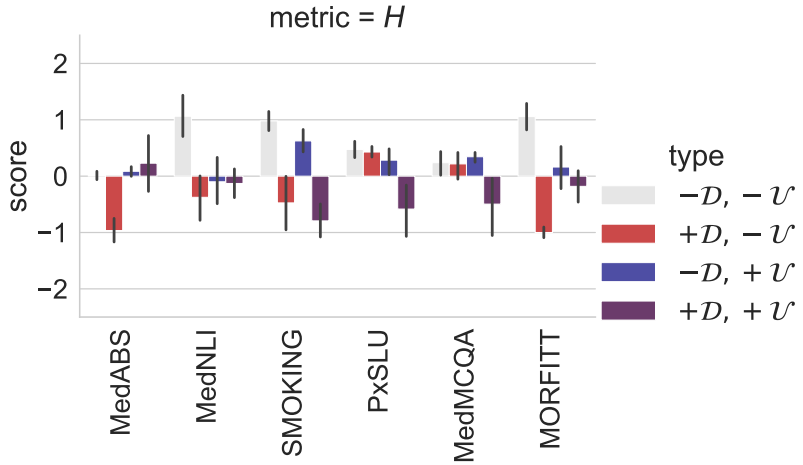



Figure 5.3: Performances on Classification metrics for empirically best models ($z$-normalized per dataset).

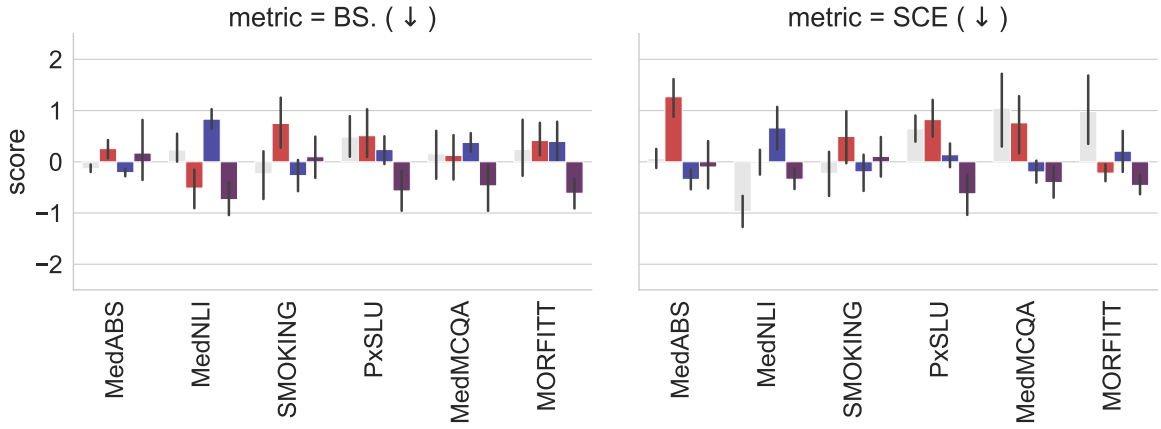
[3]Here, $z$-normalizing refers to subtracting the mean and dividing by the standard deviation of scores within each dataset. This standardization allows for fairer comparisons by removing dataset-specific biases in score magnitudes.

As for entropy, we find both $+\mathcal{D} - \mathcal{U}$ and $+\mathcal{D} + \mathcal{U}$ to lead to lower scores (see fig. 5.4a).

As for calibration metrics in fig. 5.4b, we find a very similar behavior to what we highlight in the main text: uncertainty-unaware model almost never rank among the top two contenders. Rankings per metric tend to be fairly stable as long as we control for domain-specificity. Lastly, having a look at the various Bayesian architecture (see fig. 5.5), we can see that DropConnect is not necessarily the best system across all uncertainty-aware classifiers. Selecting the best architectures given 3 seeds, and then expanding to 10 seeds most likely led to some degree of sampling bias, explaining this discrepancy. It does however constitute a strong contender across many situations: it still remains the best ranking Bayesian architecture on average both in terms of F1 across the validation set, as well as in terms of test BS., ECE, SCE, NLL and Entropy.




(a) Entropy




(b) Calibration metrics

Figure 5.4: Uncertainty quantification measures for empirically best models (selected metrics), $z$-normalized per dataset.

In fact, differences in terms of ranks across datasets per architecture are not always significant: If we normalize all 80 classifiers per dataset by taking their rank, then

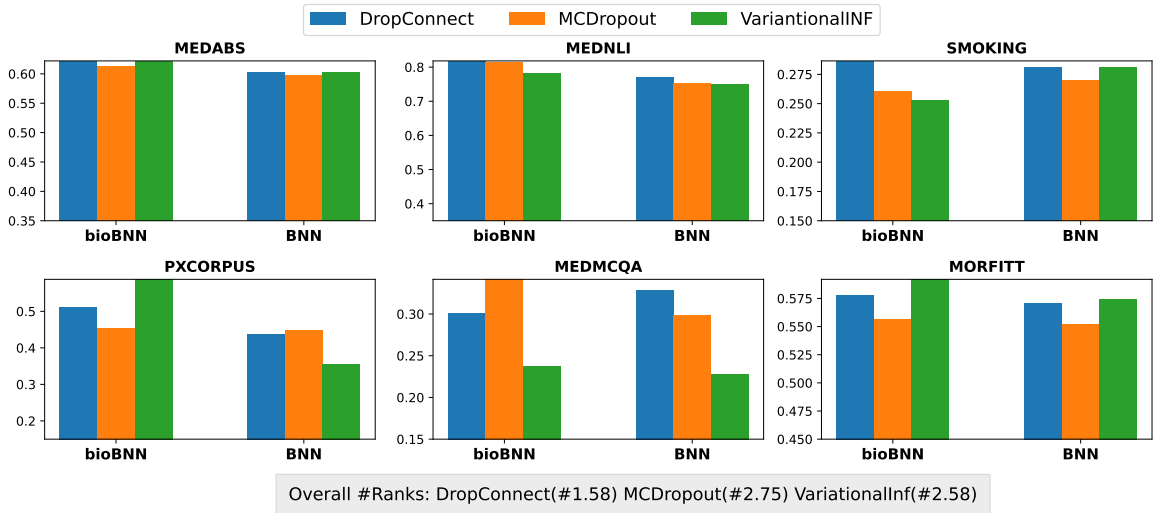



Figure 5.5: Comparison of various BNN models for different datasets on classification task based on Macro-F1 on validation set.

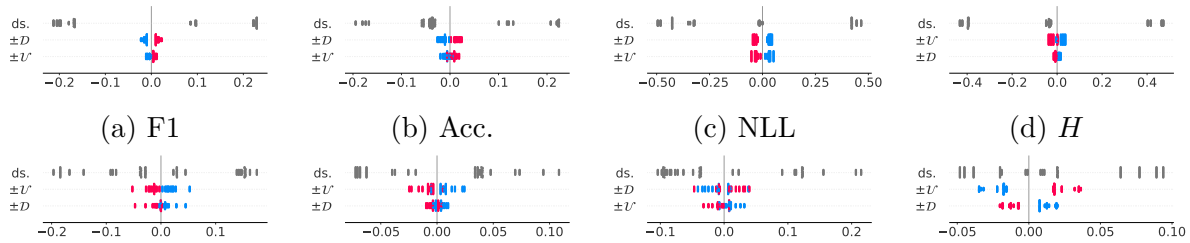
Kruskal-Wallis H-test[4] suggest that F1, accuracy and ECE do not lead to significant rank differences across architectures (assuming a threshold of $p < 0.05$). Likewise, comparing $+\mathcal{D}$ and $-\mathcal{D}$ models with the same procedure does not lead to significant differences in terms of ECE, SCE, and coverage.

One key remark is that $+\mathcal{U}$, $+\mathcal{D}$ models usually rank first both in terms of task performance and uncertainty quantification. Conversely, models of types $+\mathcal{D}$, $-\mathcal{U}$ and $-\mathcal{D}$, $+\mathcal{U}$ often yield comparable performances. Trends are consistent across languages.

Next, to interpret better the interaction between performance and uncertainty awareness and disentangle the results in figs. 5.3 and 5.4 more rigorously, we rely on SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017a). SHAP is an algorithm to compute heuristics for Shapley values (Shapley, 1953), viz. a game theoretical additive and fair distribution of a given variable to be explained across predetermined factors of interest. Here, we analyze the scores obtained by individual classifiers on all 8 metrics, and try to attribute their values ($z$-normalized per dataset) to domain specificity ($\pm\mathcal{D}$), uncertainty awareness ($\pm\mathcal{U}$) and the dataset (ds.) respectively.

Results are displayed in fig. 5.6; specific points correspond to weights assigned to one of the factors for one of the datapoints, factors are sorted from most to least impactful from top to bottom. We can see that which of domain specificity and uncertainty awareness has the strongest impact depends strictly on the metrics: Cases where $\pm\mathcal{D}$ is assigned on average a greater absolute weight than $\pm\mathcal{U}$ account for exactly half of the metrics we study. Another import trend is that effects tied to $+\mathcal{D}$ are also often attested for $+\mathcal{U}$: if domain specificity is useful, then uncertainty awareness is

[4]The Kruskal–Wallis H-test is a non-parametric statistical test used to determine whether there are statistically significant differences between the medians of three or more independent groups. It is a rank-based alternative to one-way Analysis of variance (ANOVA) and does not assume a normal distribution.




(a) F1 (b) Acc. (c) NLL (d) $H$

(e) Brier score (f) SCE (g) ECE (h) Cov %

Figure 5.6: SHAP attributions. Variables are ordered by mean absolute SHAPs. In blue, weight assigned when the variable is negative; in red, when it is positive. 'ds.' denotes a categorical variable tracking the dataset.

as well.[5] Lastly, weights assigned to both $\pm\mathcal{D}$ and $\pm\mathcal{U}$ are considerably smaller than those assigned to datasets, showcasing that these trends are often overpowered by the specifics of the task at hand.

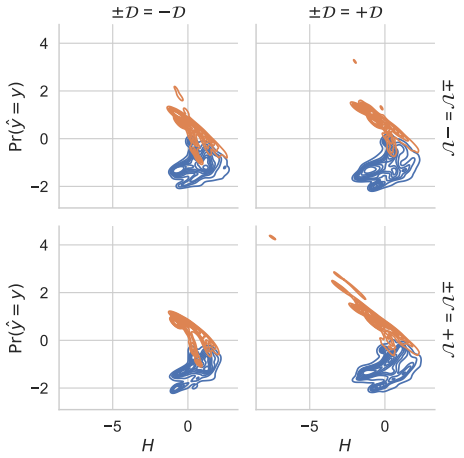



Figure 5.7: (a) Entropy vs. probability mass assigned to the target ($z$-normalized per classifier). Orange: correct predictions; Blue: incorrect.

| | effect size $f$ | | | | Spearman's $\rho$ | | | |
|---|---|---|---|---|---|---|---|---|
| | $-\mathcal{D}-\mathcal{U}$ | $+\mathcal{D}-\mathcal{U}$ | $-\mathcal{D}+\mathcal{U}$ | $+\mathcal{D}+\mathcal{U}$ | $-\mathcal{D}-\mathcal{U}$ | $+\mathcal{D}-\mathcal{U}$ | $-\mathcal{D}+\mathcal{U}$ | $+\mathcal{D}+\mathcal{U}$ |
| **MedABS** | 62.5 | 64.8 | 62.4 | **67.3** | −48.0 | −47.9 | −44.6 | **−53.5** |
| **MedNLI** | 73.2 | 73.2 | 74.0 | **77.0** | −73.2 | −77.4 | −76.1 | **−83.3** |
| **SMOKING** | **75.8** | 71.6 | 74.2 | 74.8 | **−56.5** | −38.0 | −50.0 | −56.0 |
| **PxSLU** | 65.4 | **87.2** | 65.1 | 85.8 | −85.4 | −69.1 | −87.3 | **−96.2** |
| **MedMCQA** | 65.6 | 63.8 | 66.6 | **68.2** | **−82.3** | −82.2 | −60.8 | −62.6 |
| **MORFITT** | 65.6 | **66.1** | 65.0 | 64.8 | −54.6 | **−55.1** | −50.8 | −51.0 |

Table 5.2: (b) Statistical tests on entropy measurements, with **best** and second best highlighted.

Finally, the desideratum we laid out above was to have large entropy scores when the model is incorrect. Focusing on entropy, we display how it compares to the probability mass assigned to the target in fig. 5.7. In detail, we retrieve all predictions for every datapoint across all classifiers and then $z$-normalize entropy scores and probability assigned the target class.[6]

[5] There are two notable exceptions: ECE and coverage, where we find $+\mathcal{D}$ to be *detrimental*. Variation across seeds might explain the discrepancy with B.10.

[6] When plotting entropy against probability mass assigned to the target class, we can keep in mind some useful points of reference. A perfect classifier that is always confidently correct should display a high probability mass and a low entropy (i.e., top left of our plot); what we hope to avoid is a

We can see that incorrect predictions do result in more spread out entropy scores. Moreover, we can notice some tentative differences between the four types of classifiers of our study: Correct predictions from $+\mathcal{D}+\mathcal{U}$ models seem to lead to an especially tight correlation between entropy and probability mass. However, establishing whether this difference is significant requires further testing. We therefore measure whether incorrect predictions lead to higher entropy in two ways: (i) using Mann–Whitney U-tests, from which we derive a common language effect size $f$ (as the entropy of incorrect predictions should be higher);[7] and (ii), by computing Spearman correlation coefficients between the entropy and the mass assigned to the target class (as entropy should degrade with correctness). Corresponding results are listed in Table 5.2: Across most of the datasets we study, the top or second most coherent distributions we observe are for domain-specific and uncertainty-aware models. However, we also observe that actual performances are highly sensitive to the exact classification task at hand.

We can now answer our initial research questions. (i) *Are the benefits of uncertainty-awareness and domain-specificity orthogonal?* We have seen in table 5.2 that in most cases, using a classifier that was both domain-specific and uncertainty-aware led to the optimal distribution shape, with entropy more gracefully increasing with incorrectness; (ii) *Should medical practitioners prioritize domain-specificity or uncertainty-awareness?* SHAP attributions in fig. 5.6 strongly suggest that the evaluation metric dictates the strategy to follow. As one would expect, accuracy is better captured with domain-specific models, whereas uncertainty-aware models tend to be better calibrated.

We also found significant evidence throughout our experiments that the exact classification task at hand weighs in much more strongly than the design of the classifier. This extraneous factor necessarily complicates the relationship between domain-specificity and uncertainty-awareness: In a handful of cases in figs. 5.3 and 5.4, we observe classifiers that are neither uncertainty-aware nor domain specific faring best among all the models we survey and conversely domain-specific uncertainty-aware classifiers can also rank dead last. This is also related to the often limited quantitative difference between best and worst models, which for instance can be as low as $\pm 2.3\%$ for F1 on MEDABS (cf. table B.10).

## Summary

Overall, this section points to a nuanced conclusion. While both domain-specific pretraining and uncertainty-awareness influence classifier behavior—particularly in shaping output distributions and entropy—their impact is secondary to that of the task itself. This suggests that although uncertainty-awareness and domain-specificity are often compatible and can be effectively combined, there are no universal solutions. Achieving optimal performance still requires thoughtful, task-sensitive model design.

confidently incorrect classifier (bottom left). As entropy and probability are statistically related, it is impossible to observe a high probability mass and a high entropy (top right). Lastly, assuming the classifier outputs continuous scores, this statistical dependency also dictates that probability mass and entropy be inversely correlated for correct predictions.

[7] All U-tests suggest entropy for incorrect predictions is significantly higher ($p < 10^{-10}$).

**Limitations** A key limitation of this study is the absence of large language models (LLMs) in the experimental design, which restricts the generalizability of the findings especially in the context of current trends in clinical NLP. Future work could address this by incorporating LLMs to explore how their capabilities interact with uncertainty and domain adaptation.

### 5.2.2 Interpretation of Medical Models

This chapter is based on work previously published in our article: *Sinha, A., Mickus, T., Clausel, M., Constant, M., & Coubez, X. (2025a). Simplicity isn't as simple as you think. In The 1st Workshop on Actionable Interpretability at Forty-Second International Conference on Machine Learning, Vancouver, Canada.* Parts of the text, figures, and results are adapted from this publication.

Figure 5.8: Black box Medical Models (image src: gemini-flash-2.5)

Deep Learning based models are inherently opaque, functioning as black boxes whose internal decision-making processes remain largely non-transparent. While highly effective at capturing complex patterns, their increasing depth and parameterization further obscure how input features translate into predictions. Successive layers transform data into high-dimensional representations that defy intuitive interpretation. These architectures have achieved remarkable success across domains such as computer vision, natural language processing, and healthcare, yet their predictive strength often comes at the cost of comprehensibility. This opacity raises critical concerns about reliability, hidden biases, and the risks of deploying such systems in mission-critical settings where interpretability is as vital as accuracy. The issue is particularly acute in the medical domain, where model outputs can directly influence clinical decisions and patient outcomes. Consequently, interpretability has become a central demand, enabling researchers and practitioners to examine model reasoning, align it with expert knowledge, and ensure that predictions are not only accurate but also trustworthy.

As a field of study, interpretability focuses on understanding and explaining the internal mechanisms and outputs of complex models. By providing insights into how models process input data and arrive at predictions in the form of *explanations* (can be thought of model's rationale behind the outcomes), interpretability helps identify potential biases, errors, or unexpected behaviors. Interpretability has explored several evaluation criteria that range from correctness, i.e., how closely the explanations match the model's behavior, to usefulness, i.e., how easy the explanations are for end users to understand. Prior literature consistently emphasizes correctness over usefulness, and usually discards the latter as a characteristic that requires sacrificing the former. While there are practical arguments in favor of such an approach, this state of affairs has also resulted in a knowledge gap. And therefore, in this chapter, we present an investigation with the simplicity metric (Bhatt et al., 2020) to answer the following question:

**Research Question 3.1b:** *Is simplicity a useful metric for evaluation of LMs interpretability?*

## Background

This knowledge gap led to the introduction of the *Co-12* property framework[8] (Nauta et al., 2023), which are criteria for quantitative evaluation of explanation. These criteria in summary fall into one of the following three qualitative axioms: correctness, robustness and usefulness. These axioms are well reflected in the literature via prominent methods such as faithfulness (Samek et al., 2016) as a measure of correctness and sensitivity or stability (Dai et al., 2022) as a measure of robustness and complexity or sparseness as a measure of usefulness (Chalasani et al., 2020). Most studies acknowledge the trade-off that occurs when optimizing for the three desiderata simultaneously in order to obtain end-user friendly explanations (Chen et al., 2022; Tan and Tian, 2023; Hedström et al., 2023; Bhatt et al., 2020). Among the three axioms, usefulness is the hardest to assess given that it is contextually dependent on the exact target application of a given AI system and requires human-centric assessment. Furthermore, while a consensus around the fact that we struggle to balance usefulness with correctness and robustness has emerged, the literature has yet to settle on convincing root causes for this inability.

The motivation for this study primarily stems from the expected human-centric properties of an explanation such as the semantic alignment between the user's mental model (aka. rationales) and the presentation in the explanation (also refereed to as understandability, cf. (Chen et al., 2022; Moreno-Sánchez, 2023)). For instance, classifying a chest scan as containing a tumor should ideally be done on the basis of the limited set of pixels corresponding to said tumor and its accompanying telltale signs; an equally accurate decision support system that would highlight the entirety of the image would be found less useful. Additionally, disagreement in human annotations (Sundararajan and Najmi, 2020; Atanasova, 2024) is another factor that tends to deviate the model from expected behavior (Mickus et al., 2025) which can potentially be reflected in explanations.

In the present chapter, we address this gap by exploring the interaction of human preferences and the usefulness axiom via the *simplicity* metric (Bhatt et al., 2020) which is an entropy-based indicator for evaluating the quality of model's explanation. More concretely, we study the behavior of *simplicity* metric in the context of NLP tasks where we have access to human preferences in form of rationales or multiple annotations.

[8] *Co-12* properties are Correctness, Completeness, Consistency, Continuity, Contrastivity, Covariate complexity, Compactness, Composition, Confidence, Context, Coherence and Controllability.

## Methodology

**Formulation.** We focus on low-abstraction, post-hoc interpretability methods. In an NLP setting, interpretability methods assign a real-valued importance weight $w_i$ to each token $x_i$ in a given input sentence $S = (x_1, \ldots, x_n)$ of length $n$. Formally, such a method can be represented as a function

$$I : (x_1, \ldots, x_n) \mapsto (w_1, \ldots, w_n) \in \mathbb{R}^n.$$

The magnitude $|w_i|$ reflects the estimated relevance of token $x_i$ to the model's prediction, while the sign of $w_i$ (when defined by the method) indicates whether the token contributes positively or negatively to that prediction.

A maximally simple explanation, in this setting, would simply highlight one word as relevant for the prediction at hand. As a word that is considered to weigh neither for nor against the prediction should have a corresponding weight of 0, we should assume that only one of the weights $w_i \neq 0$, and all other weights $w_1, \ldots, w_{i-1}, w_{i+1}, \ldots, w_n$ are equal to 0. Deviation away from this maximally simple explanation should be rated monotonically and smoothly higher in terms of complexity. As the interpretation functions $I$ that we consider can yield distributions of weights for sentences of arbitrary length, we would prefer any indicator not to be sensitive to sentence length.

A practical way of meeting these requirements consists in re-normalizing the weights $w_1, \ldots w_n$ into a proper distribution in the sense of probability theory, and then computing the Shannon entropy of this distribution

$$\begin{aligned} \hat{w}_i &= \frac{|w_i|}{\sum_{i=1}^{n} |w_j|} \\ \mathbb{S}_I &= \sum_{i=1}^{n} \hat{w}_i \log_n \hat{w}_i \end{aligned} \tag{5.2}$$

using the $n$ weights derived from the interpretation $I$ for a specific input sentence. Two points worth highlighting are that (i) we consider the absolute weights, i.e., we ignore sign information, when renormalizing the weights; and (ii) as a modification to Bhatt et al. (2020)'s formulation, we compute the entropy in base $n$ logarithm so as to guarantee the indicator stays between 0 and 1 for any sentence, regardless of its length $n$. Using Equation (5.2), we can verify that our maximally simple explanation yields a value of $\mathbb{S}_I = 0$, and that any deviation from this maximally simple explanation obtains a higher value.

**Dataset.** We utilize different general domain such as sentiment analysis (Stanford Sentiment Treebank (SST2); Socher et al. (2013), hate-speech (HateExplain; Mathew et al. (2021)), Natural Language Inference (NLI) from ChaosNLI (MNLI; Nie et al. (2020a); Williams et al. (2018b)) and biomedical language datasets such as medical document classification (MedABS; Schopf et al. (2023)) and medical NLI (MedNLI;

| name | TASK | Domain | train | val | test |
|---|---|---|---|---|---|
| SNLI | *inference* | Flickr30k/VisualGenome | 550152 | 10000 | 10000 |
| MNLI | *inference* | misc | 382702 | 10000 | 9815 |
| SST-2 | *single-sentence* | movie reviews | 57349 | 10000 | 872 |
| HateXplain | *single-sentence* | Twitter/Gab | 15383 | 1922 | 1924 |

Table 5.3: Dataset Description

(Romanov and Shivade, 2018)) for study the interaction between the simplicity indicator and complex phenomena setting. The statistics for the dataset are provided in Table 5.3.

## Results & Discussion

In what follows, we justify this construction empirically through human-participant case studies and model-based experimentation. We describe below three case studies we do in order to investigate the effectiveness and shortcomings of our proposed indicator in general and domain-specialized setting.

**I. Human have a preference for simplicity.** Language understanding tasks such as sentiment analysis or hate speech detection often represent the scenarios when we expect from the obtained explanations from deep learning models to be simple. More precisely, we expect specific words to be emphasized more than the other words in the sentence. This is equivalent to a human manually selecting subset of words which justifies the label. We perform a case study for SST22 (Socher et al., 2013) dataset (See Appendix B.1 for its description) where we train three classifiers for each of the 24 general domain BERT models from (Turc et al., 2019) and then apply two post-hoc methods, viz. SHAP. (Lundberg and Lee, 2017b) and Integrated Gradients (IG) (Pruthi et al., 2020), to obtain explanations.

We select an explanation pair from the pool of trained general domain BERT models, for each example in the annotation set: one from the lowest avg. $\mathbb{S}_I$ model and one from the highest avg. $\mathbb{S}_I$ model. We ask three annotators to select which of the two paired explanations fit best for each the following criteria: (i) *simplicity* (which of the explanations highlights the fewest words) (cf. compactness (Moiseev et al., 2025)), (ii) *appropriateness* (which explanation matches best with what they would have highlighted) (cf. coherence (Moiseev et al., 2025)), and (iii) *sensicality* (which of the explanations portrays the most comprehensible decision process) (cf. comprehensibility (Alangari et al., 2023)). We asked the annotators to annotate 100 samples for this study, we provided them a first round of 50 examples with the annotation guidelines followed by a discussion to clarify any confusion. The annotation guidelines can be found in the Appendix A.

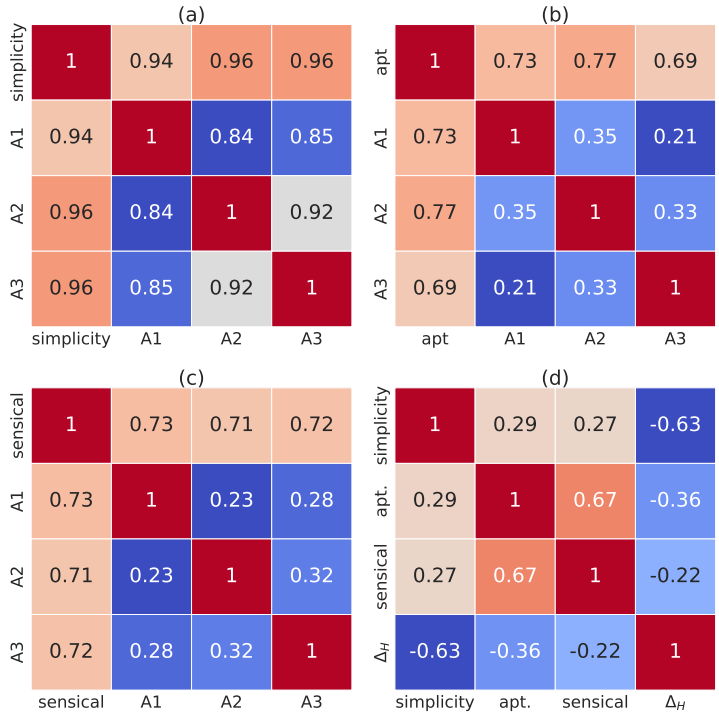



Figure 5.9: Overlap between annotator (denoted by A$i$) preference of simplicity, appropriateness (apt.), sensical and overall interaction between the three notions.

Results of the study (Figure 5.9a) show that all the annotators were able to distinguish the simpler explanation from the more complex one with a high agreement across >93% cases. Furthermore, for appropriateness, we observe an agreement of >68% (Figure 5.9b) and for sensicality, of >71% (Figure 5.9c). However, in fig. 5.9d, we observe a low spearman correlation of <30% for simplicity to appropriateness (29%) and sensicality (28%). This implies that a preference for a simpler explanation only partially translates to a preference in terms of appropriateness and sensicality for this explanation — i.e., a simple explanation tends to, but often doesn't provide sufficient attribution to relevant words in the sentence. Lastly, we remark (see Figure 5.9d) that the appropriateness of an explanation does have a strong linear correlation of 67% with its sensicality. This implies that if the explanation emphasizes the correct subset of words, it is easier to infer a reasonable prediction process as to how the task is performed.

**II. Models' interpretation are not always simple.** As in the previous setup, we train the pool of 24 BERT miniature models from (Turc et al., 2019) for hate speech detection classification, natural language inference and sentiment classification with general and domain specific datasets to investigate the generality of $\mathbb{S}_I$ across different tasks and compare the behavior for different characteristics.

**IIa. Optimal simplicity scores tend to align with human rationales** A key point worth stressing is that our proposed indicator in Equation (5.2) abstracts away the exact distribution of weights and instead focuses on a single characteristic of that distribution — while this is by design, it is still in principle possible that we find models where the indicator $\mathbb{S}_{\mathrm{I}}$ is optimal and yet the actual interpretation does not meet the expectations and preferences of human experts. It therefore stands to reason that we should assess whether or not appropriate simplicity scores correspond to adequate weight distributions.

To that end, we utilize the HateXplain dataset of (Mathew et al., 2021), which includes rationales from 3 annotators in the form of token-level attribution for sentences as a justification for the labels. For instance, for the sentence $s$ "*mudshark and black diversity*," the rationales provided by the three annotators correspond to $\mathbf{r}_{s,1} = (1, 1, 1, 0)$, $\mathbf{r}_{s,2} = (1, 0, 0, 0)$, and $\mathbf{r}_{s,3} = (1, 0, 0, 0)$, i.e., the first annotator provided the span 'mudshark and black' as a rationale, whereas the other two only highlighted 'mudshark' as an explanation. To aggregate these rationales into weight distributions comparable to what we can obtain with SHAP or integrated gradient (IG), we consider three approaches. The first considers whether any of the annotators highlighted a word, or formally, for a sentence $s$ with $n$ words:

$$\mathbf{r}_s^{\text{relax}} = \left( \bigvee_{i=1}^{3} (r_{s,i})_1, \ldots, \bigvee_{i=1}^{3} (r_{s,i})_n \right) \tag{5.3}$$

with $\bigvee$ the Boolean disjunction operator applied to its range of arguments; The second aggregation method consists in conservatively considering only the words labeled by all annotators, or

$$\mathbf{r}_s^{\text{strict}} = \left( \bigwedge_{i=1}^{3} (r_{s,i})_1, \ldots, \bigwedge_{i=1}^{3} (r_{s,i})_n \right) \tag{5.4}$$

where $\bigwedge$ is likewise the Boolean conjunction operator. The third and last aggregation method consists in simply averaging the different vectors, or:

$$\mathbf{r}_s^{\text{avg}} = \frac{1}{3} \sum_{i=1}^{3} \mathbf{r}_{s,i} \tag{5.5}$$

We can use these aggregated weight distributions to compute gold-standard simplicity scores, by first normalizing aggregated vectors as $(\hat{r}_s^{\text{agg}})_i = (r_s^{\text{agg}})_i / \sum_{j=1}^{n} (r_s^{\text{agg}})_j$ using $r^{\text{agg}}$ as a stand-in for any of the three aggregates relax, strict, or avg, and then applying Equation (5.2) to obtain an estimate of the optimal simplicity of an explanation for this datapoint, which we note $\mathbb{S}^{\text{relax}}$, $\mathbb{S}^{\text{strict}}$ or $\mathbb{S}^{\text{avg}}$. Computing these scores over the 1081 examples from the validation set for which we have rationales and where annotators all agree on at least one word, yield average values of 0.5607, 0.2383 and 0.5432 respectively. More importantly, we can now use rationales and human-derived simplicity scores to contrast explanations as yielded by SHAP or IG, to manually-coded optimal explanations, and study whether the related simplicity indicators capture the difference between the two distributions. In practice, a desired characteristic would be that a divergence between the weight distribution as encoded by humans and the

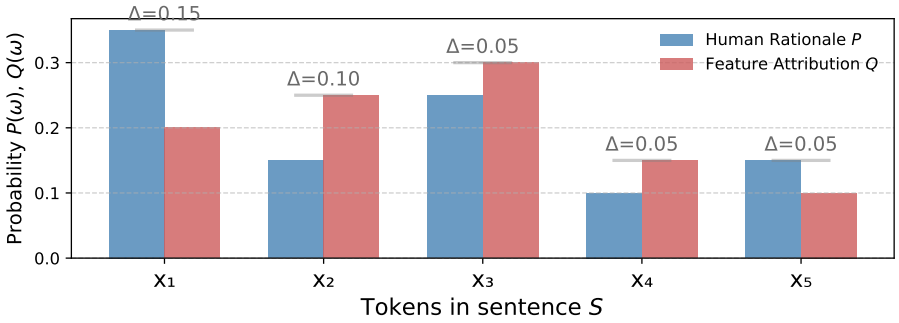



Figure 5.10: Illustration of tvd(P,Q), where P ans Q are human and model rationales.

weight distribution as estimated by an explanatory method, such as SHAP or IG, be met with a comparable difference in terms of gold and estimated simplicity indicator. That is to say, we expect the difference $|\mathbb{S}^{\text{agg}}(s) - \mathbb{S}_I(s)|$ to correlate with the divergence between the distributions $\mathtt{E}^*(\mathbf{r}_s^{\text{agg}})$ and $\mathtt{E}^*(I(s))$. For simplicity sake's, to estimate this divergence we turn to the total variation distance (see illustration in fig. 5.10), defined for two probability distributions $P$ and $Q$ over the same outcome space $\Omega$ as :

$$\text{tvd}(P, Q) = \frac{1}{2} \sum_{\omega \in \Omega} |P(\omega) - Q(\omega)| \tag{5.6}$$

where in this case, the space of outcomes is understood as the words being weighted. To establish whether a trend holds in general, we train three classifiers for each of the 24 models from (Turc et al., 2019), i.e. 72 classifiers in total. By gathering a large sample of models, we can verify which of observations hold across all models, or whether specific factors within the set of model can explain our observations. In addition to the obtained rationales from the models, we also consider 4 additional setting namely Uniform (where we sample the feature attribution weights from a uniform distribution) and then shuffled version of strict, relax and avg rationales. Summary statistics tabulated over the 72 classifiers are shown in Table 5.4 which presents the spearman correlation coefficient between the tvd and difference of the most and the least efficient language models for each of the configuration of rationales listed.

We consider as likely spurious any correlation score where the associated $p$-value is above 1/72 after a Bonferroni correction, and remove them from our analyses. Interestingly, $\mathbb{S}^{\text{relax}}$ appears the most challenging aggregation strategy to model appropriately — we observe 12 cases where the correlation is likely spurious, and effects are noticeably smaller than what we observe for other aggregation strategies. The preference for the $\mathbb{S}^{\text{relax}}$ scheme can be due to the fact that it is the one that mechanically yields Nonetheless, as is apparent, we obtain significant correlations for most of the setups we consider, underscoring that our indicator empirically lines up with an assessment of whether explanations match human rationales on HateXplain.

**IIb. Simplicity of an explanation does not relate with human label variation.** We have established that our simplicity indicator empirically correlates with an optimal explanation. A related but distinct point, is that of the uncertainty relating to a

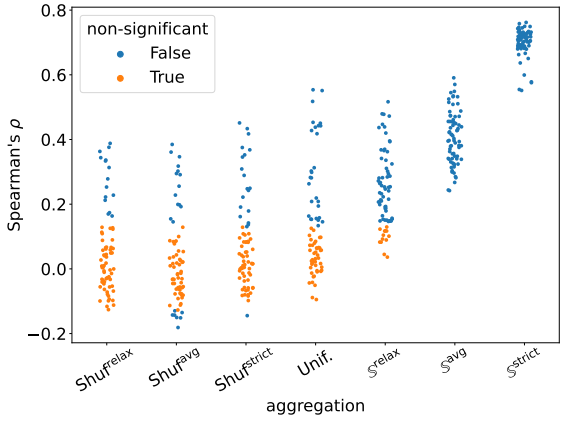



Figure 5.11: Visual Illustration of Spearman correlation between divergence and difference of simplicity for HateXplain dataset.

| Aggregate | Num. non-significant | $\rho$ when significant | | |
|---|---|---|---|---|
| | | Min | Mean | Max |
| Unif | 44/72 | 0.1336 | 0.3028 | 0.5538 |
| Shuffled strict | 50/72 | −0.1931 | 0.0975 | 0.3750 |
| Shuffled relax | 54/72 | −0.1506 | 0.2316 | 0.3890 |
| Shuffled avg | 55/72 | −0.1578 | 0.2162 | 0.3848 |
| $\mathbb{S}^{\text{relax}}$ | 12/72 | 0.1295 | 0.2697 | 0.5166 |
| $\mathbb{S}^{\text{strict}}$ | 0/72 | 0.5517 | 0.7006 | 0.7619 |
| $\mathbb{S}^{\text{avg}}$ | 0/72 | 0.2419 | 0.4061 | 0.5907 |

Table 5.4: Spearman correlation between divergence and difference of simplicity for HateXplain dataset.

sample: If for a given item, there is no clear consensus on what the label should be, then we would expect that any explanation of that item would be divided among all possible interpretations, which should mechanically increase our indicator — in other words, we expect that a greater label uncertainty correspond to a greater complexity of the corresponding explanation.

To study this aspect, we turn a large collection of human labels distributions, namely the ChaosNLI dataset of Nie et al. (2020b); for simplicity, we focus on the portions corresponding to SNLI. In details, ChaosNLI is a re-annotation of a subset of the validation splits of three NLI datasets, which contain the label judgments for 100 annotators for each datapoint, allowing researchers to look into finer-grained assessments of label uncertainty. Nie et al. (2020b) also provide a precomputed *label entropy indicator*, denoted by $\mathbb{H}_{annot}$, which comprises the label preference distribution over N classes. It is calculated as: $\mathbb{H}_{annot} = -\sum_{i=1}^{N} c_i \log_2 c_i$ where $c_i$ is the count frequency of class $i$. In practice, we compute the correlation of the entropy of human label distributions $\mathbb{H}_{\text{annot}}$ with our indicators $\mathbb{S}_{\text{SHAP}}$ and $\mathbb{S}_{\text{IG}}$, and expect to observe positive significant trend.

As before, we train classifiers for three seeds and each of the 24 models from (Turc et al., 2019), and consider as likely spurious any correlation score where the associated $p$-value is above 1/72 after a Bonferroni correction. In practice, for MNLI, none of the classifiers yield significant correlations. As for SNLI we observe that 67 out of 72 of our classifiers produce spurious correlations; with only one model consistently yielding positive correlations.[9] The five significant correlations we observe are all in the range $0.0964 \leq \rho \leq 0.1194$, that is to say, they remain noticeably small. In other words, more complex datapoints do not seem to yield more nuanced explanations.

[9] viz. `bert_uncased_L-10_H-768_A-12`; which is not the model with the highest parameter count.

| Dataset | $t$ statistic | $p$-value | Cohen's $d$ | $t$ statistic | $p$-value | Cohen's $d$ |
|---|---|---|---|---|---|---|
| | $\mathbb{S}_{\text{SHAP}}$ | | | $\mathbb{S}_{\text{IG}}$ | | |
| SST2 | $-67.1491$ | $\leq \epsilon$ | $-0.7442$ | $-67.5736$ | $\leq 10^{-70}$ | $-0.3988$ |
| HateXplain | $-18.7570$ | $\leq 10^{-70}$ | $-0.2992$ | $-26.9627$ | $\leq 10^{-70}$ | $-0.5448$ |
| MNLI | $-51.1299$ | $\leq \epsilon$ | $-0.5791$ | $-46.7358$ | $\leq 10^{-70}$ | $-0.5246$ |
| MEDABS | $-20.9117$ | $\leq 10^{-90}$ | $-0.3897$ | $18.1821$ | $\leq 10^{-70}$ | $0.3935$ |
| MEDNLI | $-21.3986$ | $\leq 10^{-88}$ | $-0.6337$ | $-38.8935$ | $\leq 10^{-224}$ | $-1.4488$ |

Table 5.5: Paired $t$-test results between the $\mathbb{S}_I$ distributions of the least and most efficient models in a pool of 72 classifiers.

**IIc. Extraneous factors can impact Simplicity.** Next, we examine whether there is any systematic trends exist across the different datasets we have considered thus far (SST2, HateXplain, SNLI and MNLI) with respect to simplcity indicator.We adopt the same approach as previously to guarantee exhaustivity, i.e., we train three classifiers for each of the 24 models from (Turc et al., 2019) and each dataset, which totals to 288 models.

First, we examine whether performance impacts the simplicity scores. We select the two models with the highest and lowest macro F1 on the validation split for any of the dataset, and then we perform $t$-test with related samples to ascertain whether the values of $\mathbb{S}_{\text{I}}$ assigned to every sample in the validation split differs between the two in a systematic fashion or not. Results are displayed in Table 5.5. Across the three datasets, we observe a very clear and significant effect regardless of the interpretation method considered. Models that perform best tend to produce explanations that are more complex: that is to say, models that are more successful also yield explanations that are less simple, as per our indicator. The only case where this pattern breaks is for MEDABS, where we observe a positive effect for $\mathbb{S}_{\text{IG}}$, i.e., the most efficient MEDABS classifier yields IG explanations that are simpler than what the least efficient MEDABS classifier yields. Effect sizes, estimated with a Cohen's $d$, are noteworthy in all cases.

One major confound we might expect is that classifiers with different F1 scores, by definition, do not perform the same predictions. In that respect, it is worth stressing that results presented in Table 5.5 still hold when we only consider items correctly classified by both models; effect sizes remain in a similar range.[10]

Yet, a broader point could be made about the relationship between the probability assigned to the target label and the value of $\mathbb{S}_{\text{SHAP}}$ we observe. For every one of our classifier, we compute the Spearman correlation between the probability mass assigned to the gold label and the corresponding $\mathbb{S}_{\text{SHAP}}$ across all validation items. As before, we consider as likely spurious any correlation score where the associated $p$-value is above $1/72$ after a Bonferroni correction, and remove them from our analyses. Corresponding results, shown in Table 5.6, demonstrate that the behavior is in this case starkly task-specific. Whereas in SST2, most of our models display a significant positive correlation (i.e., explanations are more complex when the model assigns more mass to the right

[10]With $\mathbb{S}_{\text{SHAP}}$, SST2: $d = -0.7499$; HateXplain: $d = -0.2488$; SNLI: $d = -0.6954$; MNLI: $d = -0.5706$; With $\mathbb{S}_{\text{IG}}$, SST2: $d = -0.3965$; HateXplain: $d = -0.5437$; MNLI: $d = -45.0646$; ;

| Dataset | Num. non-significant | $\rho$ when significant | | |
|---|---|---|---|---|
| | | Min | Mean | Max |
| | $\mathbb{S}_{\text{SHAP}}$ | | | |
| SST2 | 7/72 | −0.1291 | 0.2059 | 0.4567 |
| HateXplain | 20/72 | −0.2961 | −0.1583 | −0.0855 |
| MNLI | 1/72 | −0.3920 | −0.2506 | −0.0450 |
| MEDABS | 28/72 | −0.2243 | 0.0101 | 0.2440 |
| MEDNLI | 1/72 | −0.5349 | −0.3879 | −0.1298 |
| | $\mathbb{S}_{\text{IG}}$ | | | |
| SST2 | 21/72 | −0.1997 | 0.0380 | 0.2808 |
| HateXplain | 40/72 | −0.2847 | −0.1355 | 0.1958 |
| MNLI | 5/72 | −0.2913 | −0.1865 | 0.0896 |
| MEDABS | 22/72 | −0.3261 | −0.1267 | 0.1901 |
| MEDNLI | 10/72 | −0.4546 | −0.2877 | 0.1338 |

Table 5.6: Spearman correlation between probability mass on the gold label and $\mathbb{S}_I$ score across all models

| Dataset | Num. non-significant | $\rho$ when significant | | |
|---|---|---|---|---|
| | | Min | Mean | Max |
| | $\mathbb{S}_{\text{SHAP}}$ | | | |
| SST2 | 8/72 | −0.1445 | 0.2096 | 0.4692 |
| HateXplain | 13/72 | −0.4652 | −0.1955 | 0.2014 |
| MNLI | 1/72 | −0.4536 | −0.2996 | −0.0482 |
| MEDABS | 14/72 | −0.3875 | −0.0218 | 0.4280 |
| MEDNLI | 0/72 | −0.5889 | −0.4064 | −0.1039 |
| | $\mathbb{S}_{\text{IG}}$ | | | |
| SST2 | 22/72 | −0.2092 | 0.0335 | 0.2888 |
| HateXplain | 28/72 | −0.4424 | −0.1616 | 0.2626 |
| MNLI | 5/72 | −0.3580 | −0.2112 | 0.1105 |
| MEDABS | 11/72 | −0.6126 | −0.2234 | 0.3106 |
| MEDNLI | 10/72 | −0.4875 | −0.3058 | 0.1689 |

Table 5.7: Spearman correlation between probability mass on the predicted label and $\mathbb{S}_I$ score across all models

target), we see the opposite trend in SNLI and MNLI (where explanations are simpler when decisions are more straightforward); as for HateXplain , many of the correlations appear spurious, but overall somewhat in line with the NLI datasets.[11]

It is also worth stressing that considering the predicted label, instead of the gold label, does not modify results significantly, as shown in Table 5.7. In short, Tables 5.6 and 5.7 provides tentative evidence that the relationship between model distribution and simplicity of explanation is highly specific to a given scenario.

Interpretability as defined by (Madsen et al., 2022) is about explaining model's decision making process to users (Doshi-Velez and Kim, 2017), and thus this definition itself conditions the usefulness of an explanation on their relevance with respect to the preference of those who would in fact be in need of such an explanation — that is to say, explanation have to be human-grounded if they want to be useful. Our case study in section 5.2.2 demonstrate that humans can clearly distinguish between simple and complex explanations, echoing findings from previous works. This result argues in favor of the use of qualitative indicators, such as our proposed simplicity indicator, to assess the usefulness of an explanation.

| Dataset | Pearson | | Spearman | |
|---|---|---|---|---|
| | stat | $p$-value | stat | $p$-value |
| SST2 | 0.834 | 0.0000 | 0.940 | 0.0000 |
| HateXplain | 0.682 | 0.0002 | 0.676 | 0.0003 |
| SNLI | 0.795 | 0.0000 | 0.952 | 0.0000 |
| MNLI | 0.805 | 0.0000 | 0.871 | 0.0000 |
| MedABS | 0.577 | 0.0032 | 0.681 | 0.0002 |
| MedNLI | 0.794 | 0.0000 | 0.908 | 0.0000 |

Table 5.8: Linearity and monotonic correlation between performance(as a function of correct prediction count) and model complexity (as a function of number of parameters)

[11] To some extent, this idiosyncratic behavior can be explained by focusing on how we expect the tasks to be solved: for a sentiment-analysis task like SST2, we should expect that multiple lexical cues of a given polarity will strengthen a classifier's decision; whereas NLI has been documented as suffering from heuristics that models can exploit

Furthermore, the correlations we observe between simplicity and appropriateness, along with our empirical results in section 5.2.2, demonstrate that our simplicity indicator is a reasonable proxy for the selection of explanations that are more in-line with human preferences. In short, favoring simpler interpretability methods is a a reasonable rule of thumb to apply in order to guarantee explanations that more human-grounded

Our observations also suggest the existence of a tradeoff between simplicity and performance. Indeed, models with a higher performance, in the overwhelming majority of cases we cover (cf. Table 5.5), also yield more complex explanations: More sophisticated modeling techniques are able to capture more nuances of language data, which also entails that explanations of how the model processes an input need to be more nuanced as well. This also aligns with a relationship between number of parameters and model performance (cf. Table 5.8; (Kaplan et al., 2020)). is another benefit of quantitative indicators such as the one we propose. Another benefit of quantitative indicators like the one we propose is that they allow us to measure this trade-off and demonstrate that it occurs frequently but not systematically, as shown by MEDABS."

Conversely, those indicators also reveals situations where we expect a factor to be captured by the explanation, and yet one observe no such effect — in particular, in Section 5.2.2, we have remarked that explanations as produced by SHAP do not vary with respect to human label variation. If the limitations of neural networks are well documented when it comes to estimating the ambiguity inherent to human opinions, this point also indicates that our earlier remarks on the ability of a higher-performance model to capture subtler aspects of language need further contextualization: some characteristics of language remain out of reach of the default solutions we consider here.

This point is strengthen further by the observations in table 5.6: It is not the case that datapoints that models classify more confidently yield simpler explanations — the behavior here appear highly specific to the particular setup under study, with patterns ranging anywhere between strong correlations and strong anti-correlations. In a handful of cases, we can find some systematic patterns — for instance, the vast majority of NLI models yield anti-correlations, i.e., tend to yield simpler explanations for more confident decisions — but the variability we attest strongly suggest a complex relationship between decision processes and post-hoc relationships, that still needs to be accounted for. One confounding factor at hand is that post-hoc explanations need not be in line with one another. In the present case study, our working hypothesis has been to trust the validity of the explanations provided by IG and SHAP, but any practical application of our results also needs to account for their accuracy when portraying model's decision process.

Prior work on data complexity has stressed the existence of subsets of data that require a lower or higher computational effort (Gomez et al., 2022), but properly tying in how data characteristics fare with respect to the explanations produced by interpretability algorithm remain a complex, challenging task. The current difficulties we underscore are all the more concerning that the clinical setting is one where it is critical to provide reliable tools to assist practitioners and trustworthy explanations for patients that will eventually be confronted to the systems we build.

## Summary

To conclude, in this study we investigate the interaction between the simplicity indicator, an interpretability metric against human preference and language models to verify its *usefulness*.

We consider 24 models and 4 datasets to perform the study. We utilized interpretability methods, namely SHAP andIG to generate the explanations. We show that humans consistently identify simpler explanations, and simplicity tends to align with greater appropriateness of the explanation. We also uncover tentative evidence of a trade-off between performance and simplicity i.e., models that are more complex tends towards explanations that are more complex, yet we also observe that much of what explanations capture appears to be highly specific to the exact setup considered.

Simplicity indicator underscores limitations of the tools we currently rely on for interpretability a point which ought to be taken seriously if we are to eventually deploy these tools in real-world situations. However, the exact degree to which the simplest explanation is the most appropriate appears to be highly task-specific simple explanations may lack sufficient information to perform a specific task due to diverse semantic requirement of different tasks.

Overall, we argue that quantitative tools to discuss subjective explanations are an important direction to favor in future work. They provide us with means to assess what our current tools lack (such as the ability to adapt the subtlety of their explanation to the complexity of the input at hand) as well as what are the preferences of the users for whom we build these systems. Grounding explainable AI tools into human preferences using quantitative indicators is a necessary step to take if we want to build tools that are objectively reliable and trustworthy.

**Limitations** The key limitation of this work is the model diversity in experimental design. Large language models, as with very large number of parameters it would be interesting to see how much they are aligned with human's reasoning behavior.

## 5.3 Data Centric Issues

While model-centric attributes such as uncertainty and interpretability help explain part of the performance gap between medical models and human experts (Lipton, 2018; Cabitza et al., 2017), a significant portion of this gap originates from the nature of the underlying data and its perception (Ghassemi et al., 2020). In this section, we take a data-centric perspective to examine how characteristics of the training and evaluation data contribute to misalignment between model behavior and expert reasoning.

Medical (text) data is inherently complex, noisy, and often ambiguous. Clinical annotations may vary across institutions, annotators, or even within the same task—leading to significant disagreement among domain experts (Aroyo and Welty, 2015a). In such settings, expecting models to converge on a "single truth" becomes problematic (Plank et al., 2014). Instead, we must account for uncertainty and variability in both data and labels when evaluating model proficiency (Uma et al., 2021).

Moreover, human experts assess difficulty based on nuanced reasoning, prior knowledge, and contextual cues factors that may not be fully captured in training datasets. As a result, models may confidently predict incorrect answers for inherently difficult or ambiguous cases, revealing a disconnect between model certainty and actual task complexity (Finlayson et al., 2021). This raises important questions about how well models align with expert perceptions of difficulty (Baldock et al., 2021), and how disagreement manifests differently across humans and machines (Roberts et al., 2009).

In the following case studies, we explore how data-related issues, such as annotation disagreement, conceptual ambiguity, and expert-model misalignment, contribute to the proficiency gap. In particular, we focus on how models and experts differ in their understanding of clinical concepts and perception of task difficulty to overall answer the following question:

RQ 3.2

What are the data related issue that attribute to the gap?

### 5.3.1 Concept Alignment Gap

This chapter is based on work previously published in our article: *Sinha, A., Popescu, B.-V., Coubez, X., Clausel, M., & Constant, M. (2025b). Immunofomo: Are language models missing what oncologists see? arXiv preprint arXiv: 2506.11478.* Parts of the text, figures, and results are adapted from this publication.

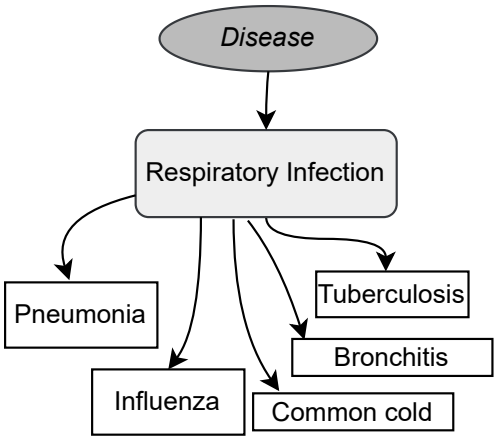



Figure 5.12: Illustration of concept structure in medical domain.

Language models (Language Model (LM)s) are designed to emulate aspects of human linguistic intelligence and have been evaluated on a wide spectrum of natural language processing (NLP) benchmarks, ranging from syntactic tasks (e.g., named entity recognition) to semantic tasks (e.g., natural language inference, document classification). In addition to their success in general NLP, LMs have also demonstrated strong performance in specialized domains such as biomedical and clinical language processing (Lewis et al., 2020a; Liu et al., 2024), as well as in complex reasoning tasks (Talmor et al., 2019; Bubeck et al., 2023).

One of the earliest domain-specific efforts in biomedical NLP was BioBERT (Lee et al., 2020), which extended BERT through pretraining on large-scale biomedical corpora. This was followed by further domain-adapted models, including SciBERT (Beltagy et al., 2019) and PubMedBERT (Gu et al., 2021a), among others. In parallel, the introduction of GPT models (Brown et al., 2020) paved the way for generative architectures, and more recently, instruction-tuned large language models (LLMs) (e.g., Zhang et al., 2023d) have been applied to clinical tasks. These models have achieved state-of-the-art results on domain-specific benchmarks such as Biomedical Language Understanding Evaluation (BLUE) (Peng et al., 2019) and Biomedical Language Understanding Benchmark (BLURB) (Gu et al., 2021a), underscoring their potential for clinical applications.

Despite these advancements, the training objectives of language models introduce fundamental limitations when applied to medical reasoning. Pretrained LMs are typically optimized for task-specific performance via supervised fine-tuning, while LLMs primarily rely on next-token prediction as a proxy for general language understanding.

Although effective, these training paradigms do not explicitly enforce the modeling of structured medical knowledge or domain-specific conceptual hierarchies.

Crucially, the nature of biomedical data itself presents additional challenges. For instance, medical knowledge often involves hierarchical and ontological relationships e.g., understanding that "pneumonia" is a subtype of "respiratory infection," which itself is a subclass of "disease" (see Figure 5.12). Such relationships are central to clinical reasoning but may not be explicitly captured by models trained on surface-level text patterns alone and contributes to the data centric issues. This raises an important research question:

**Research Question 3.2a:** *Do language models demonstrate biomedical concept understanding aligned with medical experts?*

More concretely, we present a novel evaluation dataset to study the conceptual understanding of language model against expert-labeled multi-level hallmark of cancer (in particular, Hallmarks of Immunotherapy) concepts in medical [SCI] data.

## Background

Hallmarks of cancer (Hanahan and Weinberg, 2000) have been utilized for text mining to identify the trend of cancer research focus of the disease by placing into a fixed set of alterations in cell physiology (Hanahan and Weinberg, 2000; Baker et al., 2016; Baker and Korhonen, 2017). Previous works utilized neural network and language model (Baker et al., 2016; Baker and Korhonen, 2017) for hallmark classification task and more recently, language models tailored for biomedical corpora—such as BioBERT (Lee et al., 2020), PubMedBERT (Gu et al., 2021b), and MedPalm (Singhal et al., 2025) has enabled fine-grained identification of hallmark-relevant entities, events, and pathways across large-scale literature. Despite these advances, challenges remain in capturing the implicit (linguistic) hierarchy in biomedical concepts in particular, oncology related concepts. In the context of oncology research, concepts such as hallmarks of cancer have been proposed to classify the most important principles of cancer development (Hanahan and Weinberg, 2000). The same principles of classification can also be applied to "Immune Checkpoint Inhibitors" to identify hallmarks of immunotherapy in oncology (Cogdill et al., 2017; Karasarides et al., 2022; Morad et al., 2021). Immunotherapy in breast cancer is a relatively new addition to the therapeutic arsenal (Loibl et al., 2024; Gennari et al., 2021). Pembrolizumab added to chemotherapy showed significant benefit in metastatic triple negative breast cancer in the KEYNOTE-355 trial and early triple negative breast cancer in the KEYNOTE-522 trial (Cortés et al., 2022; Schmid et al., 2020). Better understanding of such underlying hallmarks of response and resistance to immune checkpoint inhibitors in breast cancer would provide valuable data for developing novel therapeutic strategies and treatment combinations. Therefore, these hallmarks require nuanced understanding of unstructured clinical or scientific narratives. Despite this, little is known about whether existing language models can reliably recognize and reason such expert-level concepts.

Previous work in biomedical NLP has largely focused on fact-based question answer-

ing, named entity recognition, or relation extraction (Blake, 2010; Su et al., 2022). However, far fewer studies have examined the degree to which language models align with domain expert interpretations in abstract, multi-faceted concept identification tasks — particularly those that demand conceptual inference rather than surface pattern matching (Navarro et al., 2025; Workum et al., 2025). Such an evaluation benchmark tests medical models' understanding of biomedical conceptual structure as interpreted by medical experts.

Biomedical language models have shown good performance on tasks related to various medical NLP benchmarks (Yan et al., 2024). However, most of the research focuses more explicitly on encoding domain-specific semantic representation by training for sentence or document level tasks which is not directly indicative of alignment with expert conceptual interpretations. We argue that it is crucial, from a medical language understanding perspective, for language models to to have alignment with clinical experts in recognizing hierarchical concept structures (Fivez et al., 2021; Khatir and Reddy, 2024) which can foster better understanding and reliability. We therefore evaluate language models for their capability to align with expert interpretations of hierarchical biomedical concepts. More concretely, we access language models for identification of hallmarks of immunotherapy related concepts that vary from high-level to low-level , measuring agreement with clinical expert annotations. We borrow the notation of super-ordinate (high, denoted by TIER-I), basic (denoted by TIER-II) and subordinate concepts (low, denoted by TIER-III) from (Pedrotti et al., 2025).

## Methodology

**Dataset.** We extracted publicly available abstracts from multiple European Society for Medical Oncology (ESMO) congresses[12], specifically the main ESMO congress (ES), ESMO-Immunooncology (IO), and ESMO Breast (BR) conferences spanning 2020–2024, yielding approximately 10,000 abstracts in total. The dataset underwent a systematic two-stage filtering process to identify relevant studies. In the first stage (Filter1), we screened abstracts for breast cancer relevance using the keyword `breast` to capture all breast cancer-related research. Subsequently, we applied a second filter (Filter2) to isolate immunotherapy-focused studies by selecting abstracts from Filter1 that contained at least one of four predefined immunotherapy-related terms: `immunotherapy`, `immune checkpoint`, `pembrolizumab`, or `keynote`. This sequential filtering approach yielded 239 abstracts. To ensure data quality and relevance, we conducted manual verification (MANVER) of the filtered abstracts, systematically reviewing each entry to confirm its alignment with both breast cancer and immunotherapy criteria. Abstracts that did not meet these dual requirements were ex-

| | ESMO-IO | ESMO-ES | ESMO-BR | TOTAL |
|---|---|---|---|---|
| All | 810 | 8232 | 967 | 10009 |
| +Filter1 | 62 | 1646 | 967 | 2675 |
| +Filter2 | 27 | 167 | 45 | 239 |
| +MANVER | 26 | 122 | 40 | 188 |

Table 5.9: ESMO-HoI Dataset.

[12]ESMO congresses are premier global oncology meetings where leading experts present cutting-edge research, clinical advances, and therapeutic innovations in cancer care (cf. https://www.esmo.org/about-esmo-meetings/esmo-congresses).

cluded from the dataset. This final quality control step resulted in a curated dataset of 188 abstracts specifically focused on breast cancer immunotherapy research. We named our constructed dataset ESMO-HoI dataset and provide the summarized statistics in Table 5.9.

**Hallmark of Immunotherapy.** We construct an expert curated nested list of Hallmark of Immunotherapy (HoI) with help of an oncologist[13]. Each hallmark concept is then complemented with a list of sub-concepts that are associated with them. And finally, for each of the sub-concepts we create an additional list of keywords that are indicative of the sub-concepts. The primary hallmarks and their associated sub-concepts and keywords were adapted from literature and clinical experience (Morad et al., 2021; Karasarides et al., 2022; Cogdill et al., 2017). This results in a three-tier tree-like ontology where we refer to the primary hallmarks as TIER-I, the sub-concept layer as TIER-II and the keyword layer as TIER-III layer. More precisely,

- **TIER-I :** This list contains the primary nine hallmarks – "Tumor genome and epigenome", "Tumour microenvironment", "Systemic immunity", "Systemic factors", "Microbiome", "Oncogenic signaling", "Tumor metabolism", "Environmental", and "Immune checkpoint inhibitor toxicity".

- **TIER-II :** This second tier list contains 27 sub-categories. These sub-categories come from the first tier list. See table 5.10 for the mapping between tier-II and tier-I.

- **TIER-III :** This third tier list comprises relevant 177 keywords which falls into one of the 27 categories from TIER-II list. We provide table 5.11 with the mapping between tier-III and tier-I.

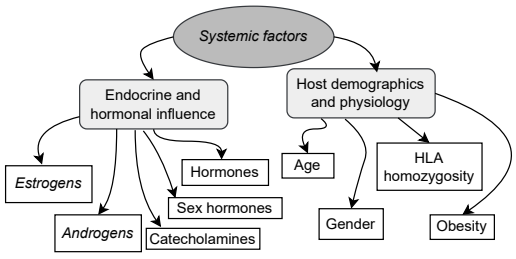



Figure 5.13: Illustration of a concept subtree from HoI labels.

[13]https://en.wikipedia.org/wiki/Oncology

| TIER-I | TIER-II |
|---|---|
| "Tumor genome and epigenome" | "Genetic alterations and immune visibility", "Epigenetic dysregulation", "Antigen presentation and immune evasion" |
| "Tumour microenvironment" | "Immune cell infiltrates", "Microbiome and extracellular factors", "Checkpoint molecules and inhibitory receptors", "Cytokines and soluble immune modulators", "Immune landscape and spatial architecture" |
| "Systemic immunity" | "Immune cell function and states", "Immune regulation and dysregulation", "Soluble mediators and modifiers" |
| "Systemic factors" | "Endocrine and hormonal influence", "Host demographics and physiology" |
| "Microbiome" | "Microbial sites", "Microbial imbalances and signatures", "Microbial species or families" |
| "Oncogenic signaling" | "Key pathways", "Genetic drivers and regulators" |
| "Tumor metabolism" | "Energy metabolism", "Metabolic reprogramming" |
| "Environmental" | "Exposures", "Psychosocial factors", "Lifestyle", "Microbial factors" |
| "Immune checkpoint inhibitor toxicity" | "Clinical manifestations", "Mechanisms", "Management" |

Table 5.10: Curated list of hallmarks of immunotherapy, in particular, a mapping between TIER-I to TIER-II concepts of hallmarks.

| TIER-I | TIER-III |
|---|---|
| "Tumor genome and epigenome" | 'CTLA-4 polymorphism', 'Human leukocyte antigen (HLA)', 'HLA heterozygosity', 'Beta-2 microglobulin', 'Tumour mutational burden (TMB)', 'Mutational load', 'Mutations/megabase', 'Mutations', 'Neoantigens', 'Genetic alterations', 'DNA mismatch repair', 'dMMR', 'Microsatellite instability (MSI)', 'Histone methyltransferases', 'Dysregulated DUX4 expression', 'Epigenetic alterations', 'DNA methyltransferase inhibitors', 'Histone deacetylase inhibitors (HDAC)', 'Epigenetic chromatin modifications', 'Histone methyltransferases', 'Epigenetic alterations', 'DNA methyltransferase inhibitors', 'Histone deacetylase inhibitors (HDAC)', 'Epigenetic chromatin modifications', 'PALB2', 'RAD51', 'Tumor mutational burden' |
| "Tumor microenvironment" | 'CD8 T cells', 'CD4 T cells', 'Tumor-infiltrating lymphocyte (TILs)', 'Teff', 'Effector T cells', 'NK cells', 'B cells', 'Tumor-infiltrating B cell', 'Tertiary lymphoid structures (TLS)', 'Treg', 'Tumor-infiltrating Tregs', 'Myeloid-Derived Suppressor Cells (MDSCs)', 'Tumour-associated macrophages (TAMs)', 'M1/M2 macrophages', 'Neutrophils', 'Cancer associated fibroblasts (CAF)', 'Fibroblast activation protein (FAP)', 'Tissue-specific stromal', 'Tumor endothelium', 'Combined Positive Score (CPS)', 'Intratumoral microbes', 'Intratumoral microbiome signature', 'Tumor microbiota', 'Tumor-associated microbes', 'Extracellular vesicles (EVs)', 'Exosomal PD-L1', 'Extracellular matrix', 'PD-1', 'PD-L1 (B7-H1)', 'PD-L2 (B7-DC)', 'CD273', 'CTLA-4', 'CD125', 'LAG-3 (CD223)', 'TIM-3', 'TIGIT', 'VISTA / VSIG', 'B7-H3', 'BTLA (CD272)', 'Siglec-15', 'VEGF', 'IL-2', 'IL-6', 'IL-10', 'IL-12', 'IL-17', 'IL-35', 'TGF-$\beta$', 'Indoleamine-2,3-dioxygenase', 'CSF1R', 'Immune excluded', 'Immune depleted', 'Angiogenesis', 'Cellular components promoting and inhibiting immunity' |
| "Systemic immunity" | 'Peripheral immune cells', 'CD8+ T-cell immune memory', 'Effector T cells', 'T-cell activation', 'T-cell exhaustion', 'Antigenic stimulation', 'Antigenic tolerance', 'Immunosuppressive phenotypes', 'Dysfunctional IFN-$\gamma$ signaling', 'Costimulation', 'Priming of immune response', 'Peripheral antigen presenting cells (APCs)', 'Host polymorphisms', 'Chronic inflammation', 'Cytokine markers', 'Soluble factors', 'Stress hormones', 'Glucocorticoids' |
| "Systemic factors" | 'Sex hormones', 'Estrogens', 'Androgens', 'Hormones', 'Catecholamines', 'Age', 'Gender', 'Obesity', 'HLA homozygosity' |
| "Microbiome" | 'Gut microbiota', 'Oral microbiota', 'Circulating microbiota', 'Dysbiosis', 'Microbiome signature', 'Bacterial 'signatures', 'Fecal microbiota transplantation', 'Bacteroidetes', 'B. fragilis', 'Bacteroides', 'Burkholderiaceae' |
| "Oncogenic signaling" | 'Interferon (IFN)', 'IFN$\gamma$ signaling', 'JAK-STAT', 'PI3K', 'MAPK', 'Wnt/$\beta$-catenin pathway', 'Hedgehog pathway', 'BRAF', 'FGFR3', 'PTEN', 'PPAR-$\gamma$', 'KRAS', 'Myc', 'BRCA', 'HER2', 'Hormone-receptor signaling', 'Estrogen-receptor (ER) signaling', 'Progesteron-receptor (PR) signaling' |
| "Tumor metabolism" | 'Glucose', 'Lactate', 'ATP', 'NAD+', 'Aerobic glycolysis', 'Oxidative phosphorylation', 'Deregulated tumor immunometabolism', 'Hypoxia', 'Oxidative stress', 'Dysregulated mitochondrial biogenesis' |
| "Environmental" | 'Food and water contamination', 'Chemical and industrial exposure', 'Household exposures', 'Air pollution', 'UV radiation', 'Pets', 'Racial injustice', 'Depression/anxiety', 'Psychological and mental stress', 'Socioeconomic inequalities', 'Sexual discrimination', 'Smoking', 'Drug use', 'Medication', 'Surgeries', 'Physical activity', 'Alcohol consumption', 'Diet and dietary supplements', 'Viral antigens', 'Pathogenic microbes', 'Household microbial contamination', 'Environmental non pathogenic microbes' |
| "Immune checkpoint inhibitor toxicity" | 'Immune toxicity', 'irAE (immune-related adverse effects)', 'Autoimmune disease', 'Aberrant T cell activity', 'Macrophage infiltrate', 'Underlying autoimmune disease', 'Use of systemic corticotherapy', 'Corticotherapy' |

Table 5.11: Curated list of hallmarks of immunotherapy, in particular, a mapping between TIER-I to TIER-III concepts of hallmarks.

**Expert Annotations.** Using the constructed HoI dataset and the curated concept taxonomy, we engaged a qualified oncologist to annotate each abstract within the HoI

dataset, assigning the most appropriate TIER-I hallmark of immunotherapy to each entry. These expert-annotated labels subsequently served as the gold standard for evaluating concept alignment in our language model analysis.

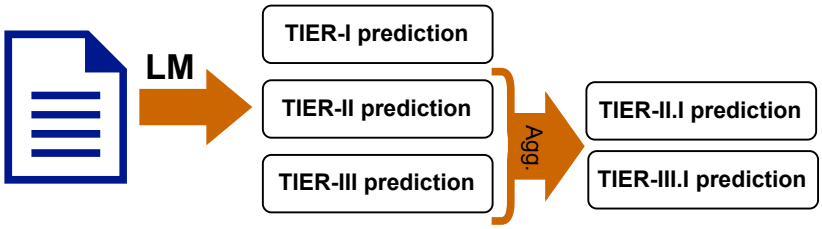



Figure 5.14: Illustration of HoI Identification task.

**Task.** We test biomedical conceptual alignment in language models via HoI identification using unsupervised text classification (Schopf et al., 2022b) as illustrated in fig. 5.14. We give as input an abstract to any language model (denoted by LM in fig. 5.14) and the task is to predict the most appropriate HoI hallmark class for each of three tiers (See Table 5.10, 5.11 for the hallmarks). Using the prediction obtained from the LM for each of the three tiers of hallmarks, we evaluate them against **TIER-I** gold labels. With **TIER-I predictions**, it is a direct comparison, however for **TIER-II, III predictions**, firstly, the obtained predictions are converted in TIER-I labels via aggregation[14] step (denoted by Agg.) to result into **TIER-II.I** and **TIER-III.I** prediction labels. And, now we compare the obtained aggregated labels against **TIER-I** gold labels (expert annotations).

**Models.** For the LM step, we consider using two approaches (a) CLS embedding based, and (b) zero shot prompting to perform the task for fair comparison across the two pool of models. In CLS based embedding approach, we utilize the [CLS] token embedding from the language model as the summary encoding of the input abstract and calculate the cosine similarity against the [CLS] token of all the different hallmark indicators for each tier. And for the zero shot prompting approach, we follow Li et al. (2023)'s technique for pre-trained language models (PLMs) i.e. Masked Language Modeling (MLM)[15] (See fig. 5.19) and we use vanilla prompting for LLMs (see fig. 5.18). We investigate two different language model categories: a collection of five small-scale pre-trained language models (PLMs) and a set of open-source large language models (LLMs). For PLMs, we employed BioBERT (Lee et al., 2020), ClinicalBERT, ClinicalBigBIRD (Li et al., 2022), SciBERT (Beltagy et al., 2019), and PubMedBERT (Gu et al., 2021a). For open-source LLMs, we utilized Llama-3-8B-Instruct (AI@Meta, 2024), Gemma-2-9B (Team et al., 2024), Med-Qwen2-7B (Bai et al., 2023), DeepSeek-R1-Distill-Llama-8B (DeepSeek-AI, 2025), and BioMistral-7B (Labrak et al., 2024).

[14]We use the term 'aggregation' for traversing up the concept hierarchy from TIER-III or TIER-II to TIER-I.

[15]In an MLM setup, the task is reformulated as a token prediction problem. For example, given the input "The color of the blood is `[MASK]`" the model fills the masked token with a suitable label word (e.g., "red"), which is then mapped to the color class *Red*.

The complete list of model configuration identifiers used in our experiments is provided in the Appendix (Table B.12).

| Metric | Criteria | Description |
|---|---|---|
| Cohen's Kappa ($C_\kappa$) | Against Oncologist | Average inter-annotator agreement between each model and the expert clinician |
| Percentage agreement (A%) | Against Oncologist | Average exact match (accuracy) between each model and the expert clinician. |
| Krippendorff's alpha ($K_\alpha$) | Against LMs | Intra-group agreement within each pool of models, treating hallucination as 'no label'. |
| Percentage consensus ($\mathbf{C}_{\delta\%}$) | Against LMs | Percentage of samples for which all models predict the same label. |

Table 5.12: Description of metrics used for evaluating alignment between LMs and experts.

**Evaluation Metrics.** We conducted a comprehensive evaluation of all language models against the oncologist-provided annotations for TIER-I labels. Since the language models also generate predictions for TIER-II and TIER-III labels, we converted these lower-level predictions to TIER-I classifications using the aggregation criteria specified in Table 5.10 and Table 5.11. Following this hierarchical aggregation, we assessed predictive performance using *weighted average F1-scores* using the expert TIER-I as gold. To evaluate concept alignment capabilities, we employed two complementary approaches: (a) agreement with expert annotations, measured through percentage agreement (A%) and Cohen's kappa ($\kappa$), and (b) intra-model pool consensus, quantified using Krippendorff's alpha ($K_\alpha$)[16] and percentage consensus ($\mathbf{C}_\delta$). Detailed descriptions of all evaluation metrics are provided in Table 5.12.

[16]We employ Krippendorff's alpha rather than Fleiss' kappa to accommodate instances where LLMs produce hallucinated or invalid responses, which we categorize as non-predictions.

# Results & Discussion

| | CLS Embedding | | | LM Prompting | | |
|---|---|---|---|---|---|---|
| | **TIER-I** | **TIER-II** | **TIER-III** | **TIER-I** | **TIER-II** | **TIER-III** |
| Clinical-BigBIRD | 0.2131 | 0.1512 | 0.1879 | 0.0005 | 0.0152 | 0.1935 |
| PubmedBERT | **0.329** | **0.431** | **0.497** | 0.0245 | 0.1470 | 0.1499 |
| BioBERT | 0.2104 | 0.2009 | 0.3266 | 0.0706 | 0.1225 | 0.2238 |
| BioClinicalBERT | 0.1983 | 0.1527 | 0.2309 | 0.0566 | 0.1109 | 0.1505 |
| SciBERT | 0.2495 | 0.1527 | 0.2257 | 0.1205 | 0.1591 | 0.1591 |
| Llama-8B | 0.2476 | 0.2909 | 0.2713 | 0.3785 | 0.3812 | 0.2694 |
| Gemma-9B | 0.1111 | 0.2359 | 0.4162 | **0.445** | **0.472** | **0.381** |
| MedQwen-7B | 0.2349 | 0.1740 | 0.0807 | 0.2468 | 0.1967 | 0.2806 |
| DeepSeek-R1 | 0.2207 | 0.3839 | 0.4622 | 0.4207 | 0.3880 | 0.2468 |
| BioMistral-7B | 0.1473 | 0.3211 | 0.3384 | 0.2163 | 0.2570 | 0.1298 |

Table 5.13: Weighted F1 score comparison between PLMs and LLMs under two different approaches

We present a comparative analysis of pre-trained language models and large language models against expert annotations for HoI classification with detailed scores provided in Table 5.13. Our evaluation encompasses two distinct methodological approaches: CLS embedding-based and zero shot prompting-based classification. For the CLS embedding-based approach, PLMs consistently outperform LLMs across all three tiers, with PubMedBERT achieving the highest weighted average F1-scores at every tier level. Conversely, for the prompting-based approach, LLMs demonstrate superior performance compared to PLMs, with Gemma-2-9B emerging as the top-performing model across all evaluation criteria. For subsequent analysis, we adopt the methodologically optimal approach for each model category: CLS embedding-based results for PLMs and prompting-based results for LLMs, ensuring fair comparison by leveraging each model type's inherent strengths.

We observed a overall low performance (range between 0.05 - 49.7% Weighted F1 Score; see Table 5.13) for PLMs which is intuitive as they are traditionally trained in a supervised/semi-supervised setting and then tested on downstream tasks. However, even with no supervised training, PubmedBERT and BioBERT performed quite comparable against LLMs especially for low-level hallmarks confirming their rich biomedical representation capacity. For LLMs, we notice some unexpected behavior as MedQwen-7B and BioMistral-7B were outperformed by general domain LLMs. LLMs did not demonstrate success on HoI classification task due to their struggle over TIER-III label identification.

| | **TIER-I** | | | **TIER-II** | | | **TIER-III** | | |
|---|---|---|---|---|---|---|---|---|---|
| | Inter-annotator (Against Expert) | | | | | | | | |
| | A% | | $\kappa$ | A% | | $\kappa$ | A% | | $\kappa$ |
| PLMs | 31.70 | | 0.11 | 32.45 | | 0.062 | 32.66 | | 0.101 |
| LLMs | 37.02 | | 0.178 | 34.260 | | 0.179 | 34.04 | | 0.169 |
| | Intra-annotator (Against Models) | | | | | | | | |
| | $\mathcal{C}_{\delta=1}$ | $\mathcal{C}_{\delta=0.8}$ | $\mathcal{C}_{\delta=0.6}$ | $\mathcal{C}_{\delta=1}$ | $\mathcal{C}_{\delta=0.8}$ | $\mathcal{C}_{\delta=0.6}$ | $\mathcal{C}_{\delta=1}$ | $\mathcal{C}_{\delta=0.8}$ | $\mathcal{C}_{\delta=0.6}$ |
| PLMs | 17.55 | 36.7 | 71.28 | 0 | 31.91 | 69.68 | 0 | 13.30 | 46.28 |
| LLMs | 10.11 | 29.79 | 62.23 | 3.19 | 17.02 | 52.66 | 2.66 | 18.62 | 67.55 |

Table 5.14: Alignment of LMs for HoI concepts.

Next, we assess the alignment between each language model and the oncologist using accuracy and Cohen's kappa metrics for inter-annotator agreement. Table 5.14 presents the alignment results for both PLM and LLM pools against clinical expert annotations. We observe overall low agreement between LMs and the clinical expert across all hierarchical levels ($\kappa < 0.20$). LLMs demonstrate low yet better agreement than PLMs across tiers. PLMs showed a drop of 5% in overall agreement at TIER-II, whereas LLMs were mostly consistent expect a slight drop for TIER-III. We additionally utilize the HoI predicted class distribution from all LMs and compare it with medical expert preference using the Jensen-Shannon distance (See fig. 5.15). We observe although overall LLMs are closer to human medical expert however, Pubmed-BERT, a PLM model is even closer compared to all the LLMs for Tier-II and Tier III HoI predictions.

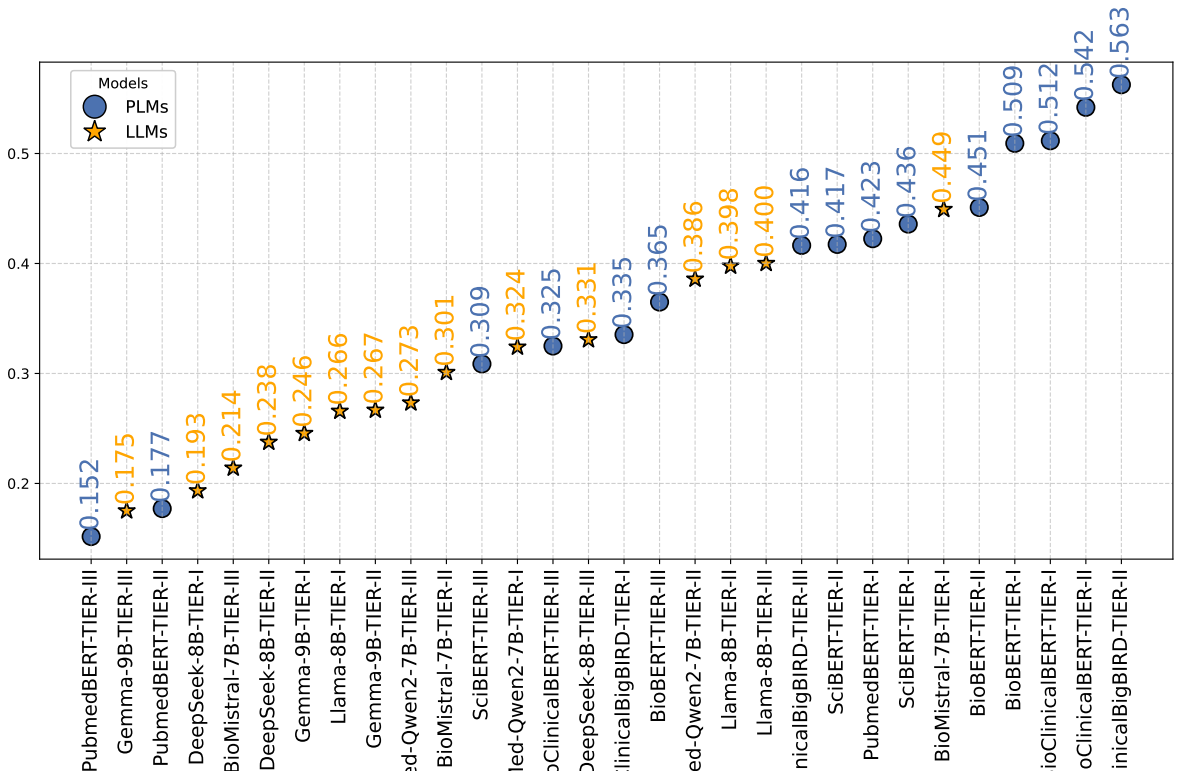
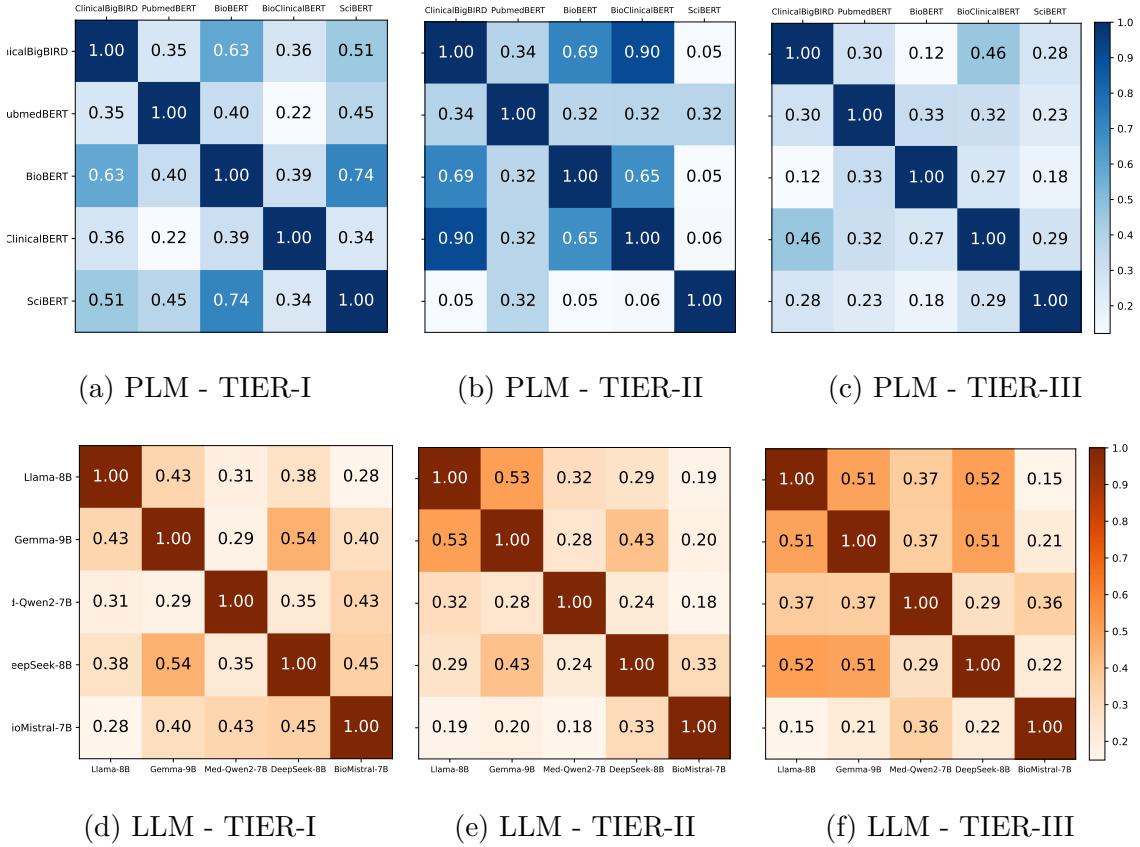



Figure 5.15: Jensen-Shannon distances comparison between PLMs and LLMs.

Table 5.14 reports the overall intra-model agreement with consensus percentage for PLMs and LLMs. Figure 5.16 presents the agreement confusion matrix for every pair of language models. In the first row (a-c), we show the confusion matrix for PLM pool with ordering of TIER-I to TIER-III from left to right, and in the second row (d-f), we have the same for the pool of LLMs. We observe that there is very low consensus ($\mathbf{C}_{\delta=1} < 4\%$) in all LMs for TIER-II and TIER-III hallmark concepts when threshold is 100%[17] however when the threshold is relaxed to 80% or lower we notice a decreasing trend of consensus with respect to concept tier granularity for all LMs. For lower concept tiers, we notice different trend for the two pools of models, for LLMs there is a consistent yet steep drop of consensus from TIER-II to TIER-III; whereas for PLMs which notice an elbow like trend across the concept tiers. The mutual alignment between LMs (cf. Figure 5.16) indicate that PLMs have better alignment for TIER-II and LLMs are more aligned for TIER-III. We notice that domain specific LLMs are aligned with general domain only TIER-I concepts.

[17]We use threshold $\delta$ to perform both strict and relaxed consensus evaluation between language model pool. With a threshold of $\delta$=1 i.e. 100% consensus, we report the what percentage of the predicted label were agreed by all the models in the pool.




(a) PLM - TIER-I (b) PLM - TIER-II (c) PLM - TIER-III

(d) LLM - TIER-I (e) LLM - TIER-II (f) LLM - TIER-III

Figure 5.16: Detailed Intra-alignment (A%) between LMs over three tiers of HOI concepts (left to right).

We perform additional investigation of the predictions obtained from the LM Prompting based approach for LLMs as they are prone to hallucinations. Table 5.13 reports Strict evaluation where we flag any predicted HoI concept that did not existed in the expert curated list (cf. table 5.10, table 5.11). For relaxed evaluation, we post-process the obtained predictions from LLMs and convert them into the closest possible HoI concept[18] corresponding to the TIER predictions. Table 5.15 report the change in hallucination percentage before and after the post-processing and additionally the resultant change in HoI identification performance. Both for TIER-I and II, LLMs suffered relatively less hallucination across all the models with the exception of BioMistral-7B which contained the highest percentage of

| | TIER-I | | TIER-II | | TIER-III | |
|---|---|---|---|---|---|---|
| Llama-8B | 0 | 0 | 0.5 | 0.5 | 56.9 | 4.8 |
| | 0.378 | 0.378 | 0.381 | 0.381 | 0.269 | 0.332 |
| Gemma-9B | 1.1 | 0.5 | 1.6 | 1.1 | 48.9 | 13.8 |
| | 0.445 | 0.450 | 0.472 | 0.477 | 0.381 | 0.544 |
| MedQwen-7B | 1.1 | 1.1 | 4.2 | 3.7 | 50 | 37.2 |
| | 0.247 | 0.247 | 0.197 | 0.202 | 0.281 | 0.328 |
| DeepSeek-R1 | 0 | 0 | 4.8 | 4.2 | 37.2 | 15.9 |
| | 0.421 | 0.421 | 0.388 | 0.387 | 0.247 | 0.335 |
| BioMistral-7B | 0.5 | 0.5 | 27.6 | 27.2 | 75.5 | 66.5 |
| | 0.216 | 0.216 | 0.257 | 0.265 | 0.130 | 0.153 |

Table 5.15: Hallucination percentage and improvement for LLMs under LM prompting evaluation.

[18]This step was done by discussion with medical expert in order to incorporate expertise while analyzing the predictions.

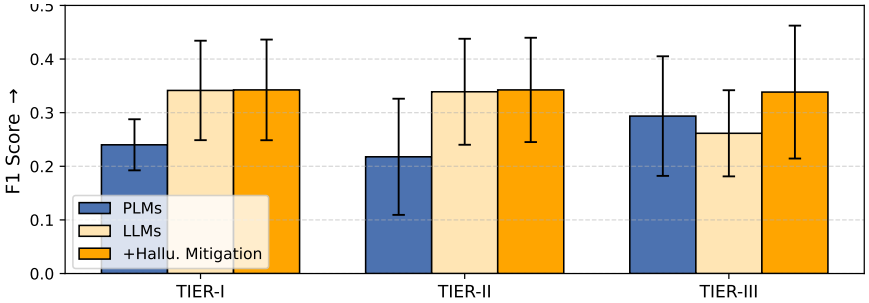



Figure 5.17: Impact of Hallucination Mitigation on LLMs performance

hallucinated output. TIER-III level was observed to be the most difficult for all LLMs with the range of hallucination in predictions to be between 37%-75%. We then performed hallucination mitigation by doing manual analysis of the hallucination that occurred during prediction step and followed by post processing of all the flagged instances during the strict evaluation step to allocate it with the most appropriate valid concept from our expert curated list. Besides showing the better capability for understanding high-level concepts, large language models tend to produce unwanted or at times incorrect information due to having access to tremendous amount of information sources which can be pre-dominately generic. The higher percentage of hallucinations in the case of TIER-III (low-level) implies that providing an explicit list of potential classes doesn't restrict the LLM to hallmark labels of interest. It was observed that when LLMs are asked to identify low-level HoI (i.e., TIER-II and III) for input, they hallucinate producing TIER-II or TIER-I related hallmarks or more generic outputs such as 'immunotherapy' or 'breast cancer'. However, the post-processing of LLM prediction led to major improvement in their performance for higher level of concepts as they even outperformed PLMs for TIER-III concepts.

## Summary

To conclude, in this case study we evaluated two diverse sets of language models for their alignment with medical expert understanding of biomedical concepts. We observe that current language models are still far from achieving strong alignment with human medical experts. While large language models (LLMs) generally achieve better overall alignment, the competitive performance of PubMedBERT, a domain-specific pretrained language model, and its outperformance of nearly all other models except Gemma-2-9B, suggests that it may be unwise to entirely abandon domain-specific PLMs in favor of the current trend toward general-purpose LLMs. Hallucination mitigation additionally arises to be an important factor to be taken into account while working with LLMs for expert domain.

**Limitations** One limitation of this study is the absence of expert annotation at intermediate concept levels (TIER-II and TIER-III). In the absence of such annotations, we evaluated the level of agreement between models as a proxy for alignment with human

expert reasoning. Intermediate-level annotations would enable a deeper understanding of how models map fine-grained concepts to abstract categories. However, the large number of labels in TIER-III (low-level) makes it difficult for models to accurately infer TIER-I annotations.

Another limitation is the filtering criteria used to select abstracts. The dataset was filtered using only four keywords: *"immunotherapy"*, *"immune checkpoint"*, *"pembrolizumab"*, and *"keynote"*. Expanding this list to include additional relevant terms could increase the size and diversity of the dataset, leading to more robust analyses of immunotherapy strategies in breast cancer.

A further limitation concerns the coverage of TIER-III keywords. The list can be extended by incorporating additional domain-specific concepts that were not initially considered. Doing so could reduce hallucinations by LLMs, as more specific and relevant terms would be available for accurate interpretation. Moreover, some keywords could be reclassified under multiple TIER-I hallmarks to reflect their relevance across different conceptual categories.

Finally, the lack of agreement between model predictions (see table 5.14) may mirror disagreements that would also occur among human annotators (Fu et al., 2024). Including a larger number of expert annotators in the evaluation process could help contextualize model disagreement and offer better insights into the variability of expert judgment.

Your task is to classify the given abstract based on the following 9 categories :
[ "Tumor genome and epigenome",
"Tumour microenvironment",
"Systemic immunity",
"Systemic factors",
"Microbiome",
"Oncogenic signaling",
"Tumor metabolism",
"Environmental",
"Immune checkpoint inhibitor toxicity"
]
you should return the top three suited categories (in top to bottom order) accurately as a python list. Abstract:

Durvalumab could be effective in combination with anti-HER2 agents in HER2-low breast cancer (ID 181)
The clinical challenge for treating HER2 (human epidermal growth factor receptor 2)-low breast cancer is the paucity of actionable drug targets. However, the discovery of immune checkpoint inhibitors has made immunotherapy an emerging new treatment modality for breast cancer. Moreover, several chemotherapeutic agents are known to induce immunogenic cell death by activating the immune system. Therefore, we hypothesized that modulating the tumour microenvironment using trastuzumab and or trastuzumab deruxtecan (T-Dxd) in breast organoids co-cultured with T-cells might enhance the response to immunotherapy. We established a panel of HER2-low breast cancer patient-derived organoids (PDOs), recapitulating the derived tumour. These PDOs were cocultured with immune cells (T- cells and Natural killer cells (NK cells)) and treated with T-Dxd and or trastuzumab in combination with durvalumab. Levels of cytotoxic markers were assessed using flow cytometry and cytokine assays. Our findings revealed synergistic effects in HER2-low BC patient-derived organoids when treated with T-Dxd and or trastuzumab in combination with durvalumab. We also observed antibody-dependent cellular cytotoxicity (ADCC) response with trastuzumab in combination with durvalumab. These results highlight the need to develop a combination treatment of PD-1/PD-L1 inhibitors with targeted therapies, and other immunotherapies to maximize clinical efficacy. Altogether, despite preliminary, these findings support the rationale for combining anti-HER2 therapies with immunotherapy in HER2-low BC patients.

Figure 5.18: Zero-shot prompting template used for LLM based HOI identification.

Text :
Durvalumab could be effective in combination with anti-HER2 agents in HER2-low breast cancer (ID 181)
The clinical challenge for treating HER2 (human epidermal growth factor receptor 2)-low breast cancer is the paucity of actionable drug targets. However, the discovery of immune checkpoint inhibitors has made immunotherapy an emerging new treatment modality for breast cancer. Moreover, several chemotherapeutic agents are known to induce immunogenic cell death by activating the immune system. Therefore, we hypothesized that modulating the tumour microenvironment using trastuzumab and or trastuzumab deruxtecan (T-Dxd) in breast organoids co-cultured with T-cells might enhance the response to immunotherapy. We established a panel of HER2-low breast cancer patient-derived organoids (PDOs), recapitulating the derived tumour. These PDOs were cocultured with immune cells (T- cells and Natural killer cells (NK cells)) and treated with T-Dxd and or trastuzumab in combination with durvalumab. Levels of cytotoxic markers were assessed using flow cytometry and cytokine assays. Our findings revealed synergistic effects in HER2-low BC patient-derived organoids when treated with T-Dxd and or trastuzumab in combination with durvalumab. We also observed antibody-dependent cellular cytotoxicity (ADCC) response with trastuzumab in combination with durvalumab. These results highlight the need to develop a combination treatment of PD-1/PD-L1 inhibitors with targeted therapies, and other immunotherapies to maximize clinical efficacy. Altogether, despite preliminary, these findings support the rationale for combining anti-HER2 therapies with immunotherapy in HER2-low BC patients.

is this about «Tumor genome and epigenome»? [MASK]

Figure 5.19: Masked Language prompting template used for PLM based HOI identification.

### 5.3.2 Difficulty Perception Gap

This chapter is based on work previously published in our article: *Mickus, T., Sinha, A., & Vázquez, R. (2025). Your model is overconfident, and other lies we tell ourselves. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5401–5417, Vienna, Austria. Association for Computational Linguistics.* Parts of the text, figures, and results are adapted from this publication.


Figure 5.20: The Rabbit–duck illusion

A central, yet often underappreciated, aspect of natural language processing is the dissensus among annotators when labeling a given datapoint. Human annotators frequently provide differing labels, and prior research shows that such disagreements are not merely the result of random noise but instead reflect intrinsic properties of language and meaning (Plank et al., 2014; Plank, 2022; Uma et al., 2021). This *annotator dissensus* is closely linked to *data complexity*: examples that challenge human judgment tend to be more difficult for models as well. Data complexity can be measured through training dynamics (e.g., how quickly a model learns an example), model confidence, or variability in model performance across multiple runs or architectures (Guo et al., 2017; Swayamdipta et al., 2020; Hendrycks and Dietterich, 2019). Understanding this relationship has important implications for both model development and the construction of evaluation benchmarks, which increasingly seek to capture human uncertainty through multiple annotations per datapoint (Bowman et al., 2015; Nie et al., 2020b).

Several factors contribute to data uncertainty, including annotation noise, semantic ambiguity, and overlapping class boundaries (Hu et al., 2023). These factors not only lower inter-annotator agreement but also highlight genuine complexity in language and reasoning, a phenomenon observed across a wide range of NLP tasks (Bowman et al., 2015). While such patterns are particularly consequential in high-stakes domains like medicine—where the gap between model predictions and expert judgment can have critical consequences—they are equally relevant in general-domain tasks, providing a controlled setting to systematically study how models and humans respond to complex data.

This leads to a fundamental question at the intersection of human cognition and model learning:

**Research Question 3.2b:** *Is data complexity same for both LMs and human experts?*

## Background

Our study falls at the intersection of data complexity and uncertainty in NLP, particularly in relation to annotator disagreement, label variation, and the metrics used to evaluate them (Jiang and de Marneffe, 2022; Uma et al., 2021; Lorena et al., 2019; Baan et al., 2023). Though closely related, these concepts involve distinct challenges. **Data uncertainty** or aleatoric uncertainty describes the *randomness or noise inherent to the data* (Hu et al., 2023). This type of uncertainty is *irreducible* and cannot be eliminated through model improvements or tuning (Kiureghian and Ditlevsen, 2009; Kendall and Gal, 2017; Hüllermeier and Waegeman, 2021). Sources of data uncertainty include noisy observations, overlapping classes, ground truth errors or inherent randomness. **Annotator disagreement** is highlighted as *a fundamental characteristic of linguistic data*, stemming from both annotation noise and the inherent ambiguity of language (Plank et al., 2014; Aroyo and Welty, 2015b; Pavlick and Kwiatkowski, 2019; Fornaciari et al., 2021). It often correlates with *label uncertainty*, where no single correct label exists. High-disagreement examples contain valuable signals for classifiers (Basile et al., 2021; Palomaki et al., 2018). However, disagreement does not necessarily indicate annotation noise—it may instead reflect genuine linguistic or contextual ambiguity. **Data complexity** or data difficulty refers to the *characteristics of a data sample that make classification inherently difficult*. It is related to the structural properties of the data, not to randomness or noise, as is the case with data uncertainty. Several factors contribute to data complexity: proximity to decision boundaries and class overlap, semantic indeterminacy, and task-specific challenges, such as requiring world knowledge (Plank, 2022; Jiang and de Marneffe, 2022). **Metrics for evaluating data uncertainty and complexity** include model confidence (probability of the predicted class), entropy of predicted probabilities (measuring classification uncertainty), and confidence calibration (aligning confidence with performance). These help assess label uncertainty caused by overlapping class boundaries e.g., (Geng et al., 2024; Zhou et al., 2022; Xiao and Wang, 2019b). Training dynamics, such as how quickly a model learns to classify an example or the shape of the loss curve, further reveal the relative difficulty of datapoints (Swayamdipta et al., 2020; Toneva et al., 2019; Baldock et al., 2021). These metrics offer tools to analyze the challenges of data complexity, yet they do not provide a nuanced perspective on the interplay between annotator disagreement and model uncertainty. Data complexity, annotator disagreement, and data uncertainty are intertwined phenomena with similar root causes. For instance, while authors differ in their terminology, 'overlapping classes' (Ho and Basu, 2002; Peterson et al., 2019; Lorena et al., 2019), 'absence of a single ground truth' (Aroyo and Welty, 2015b; Baan et al., 2022), or 'linguistic ambiguity' due to semantic and social factors (Plank, 2022) all refer to the fact that some datapoints can have different labels — which drives up complexity, disagreement and uncertainty.

Hence, these three phenomena have been conflated in the literature. For example, uncertainty is often measured through label distribution entropy, used as a proxy for

| Human-Based | Model-Based | |
|---|---|---|
| | `Reference-free` | `Reference-dependent` |
| Empirical population dissensus | Model pool dissensus | Model pool failure rate |
| Empirical population entropy | Model pool entropy | Early computation termination |
| | Averaged model entropy | Early training termination |
| | Conformal prediction set size | Failure rate through training |
| | | Probability mass through training |

Table 5.16: Our taxonomy for the different difficulty indicators that are used in this case study.

annotator disagreement (Zhang et al., 2022; Baumler et al., 2023). Similarly, (Lalor et al., 2018) has found alignment between human difficulty and model-assigned probability mass, suggesting that both perceive difficulty similarly. This viewpoint is also prevalent in active learning e.g.,(Hachey et al., 2005) and attested less directly in quality estimation e.g., (Jamison and Gurevych, 2015). Additionally, anecdotal evidence has been used to support this connection, especially in studies exploring the underlying causes of data complexity (Swayamdipta et al., 2020; Baldock et al., 2021; Rajpurkar et al., 2018). This perspective has also been employed as a working hypothesis. For example, (Weinshall et al., 2018) assume that knowledge distillation from a model can play the same role as humans in active learning scenarios, while (Beigman and Beigman Klebanov, 2009) propose replacing annotator-based disagreement assessment with classifier proxies. Authors adopting this approach often treat it as a simplifying initial assumption, and frequently include modeling work to better capture human variation e.g., (Reidsma and Carletta, 2008) or discussions of the limitations of this working hypothesis (as in (Beigman and Beigman Klebanov, 2009)). Treating these three phenomena as interchangeable oversimplifies their relationships. Disagreement often signals semantic complexity but can also stem from bias, expertise variance, or cultural differences e.g.,(Jiang and de Marneffe, 2022). Similarly, uncertainty overlaps with complexity but also arises from noise that is not tied to structural data complexity (Kendall and Gal, 2017), while example difficulty has been linked to factors that are *a priori* not linguistic, such as class imbalance or distributional shifts (Ho and Basu, 2002; Gawlikowski et al., 2023). While some studies find overlap between human disagreement and model uncertainty (Swayamdipta et al., 2020; Baldock et al., 2021), others challenge this view (Reidsma and Carletta, 2008). Our findings highlight the need to distinguish these concepts, as we show that complexity metrics do not map linearly to human assessments. We formalize several means of assessing how difficult it is to assign one of $k$ possible labels $\{y_1, \ldots, y_k\} = Y$ to a specific instance $x$ (See the list in Table 5.16) and describe them formally below:

1. **Empirical population dissensus.** The simplest way to quantify disagreement on a specific datapoint is to ask multiple annotators $a_1, \ldots, a_n$, and compute how unpopular the majority opinion is. As such, if annotator $a_j$ would assign the label $y_{a_j}$ to the observation $x$, we can define the probability $\Pr_{\mathbb{H}}(y_i|x)$ on label $y_i$

as the proportion of annotators agreeing on the label $y_i$, or formally

$$\Pr_{\mathbb{H}}(y_i|x) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\left\{y_{a_j} = y_i\right\}$$

and denote the dissensus among annotators as:

$$\mathbb{H}_{\text{dis}} = 1 - \max_{y_i \in Y} \Pr_{\mathbb{H}}(y_i|x), \tag{5.7}$$

$\mathbb{H}_{\text{dis}}$ is hence inversely related to the popularity of the most common label. If all annotators agree, there is a strong consensus, implying that $\mathbb{H}_{\text{dis}} = 0$ because $\max_i \Pr_{\mathbb{H}}(y_i|x) = 1$.

2. **Empirical population entropy.** The empirical dissensus $\mathbb{H}_{\text{dis}}$ has the drawback of not factoring in minority opinions: there is a distinction to be made between having the opinions split among a handful of well-supported alternatives versus a total lack of consensus and annotators maximally split across all possible alternatives. To account for such differences, we consider the empirical entropy of opinions (Nie et al., 2020b), or

$$\mathbb{H}_{\text{ent}} = -\sum_{y_i \in Y} \Pr_{\mathbb{H}}(y_i|x) \log \Pr_{\mathbb{H}}(y_i|x) \tag{5.8}$$

Entropy measures uncertainty or diversity in the label distribution, better accounting for both dominant and minority labels. As before, $\mathbb{H}_{\text{ent}} = 0$ when all annotators agree on one label. Contrastingly, $\mathbb{H}_{\text{dis}}$ is maximal when $\Pr_{\mathbb{H}} \sim \text{Unif}$, i.e., when annotators are evenly split across all labels.

3. **Model pool dissensus and model pool entropy.** Given a set of models parametrized by $\theta_1, \ldots, \theta_m$, we can easily extend the concepts of dissensus ($\mathbb{H}_{\text{dis}}$) and entropy ($\mathbb{H}_{\text{ent}}$) to models' predictions, instead of relying on human annotators. To do this, we evaluate the predictions of a model $\theta_j$ by selecting the label $\arg\max_{y_i \in Y}\ p(y_i|x, \theta_j)$. Next, we define the probability $\Pr_{\mathbb{M}}(y_i|x)$ of this datapoint being labeled as $y_i$, by tallying the number of models that predict $y_i$ as the most likely label:

$$\Pr_{\mathbb{M}}(y_i|x) = \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\left\{y_i = \arg\max_{y_k \in Y}\ p(y_k|x, \theta_j)\right\}$$

Using this distribution, we can analogously define both metrics for the models' predictions:

$$\mathbb{M}_{\text{dis}} = 1 - \max_{y_i \in Y} \Pr_{\mathbb{M}}(y_i|x) \tag{5.9}$$

$$\mathbb{M}_{\text{ent}} = -\sum_{y_i \in Y} \Pr_{\mathbb{M}}(y_i|x) \log \Pr_{\mathbb{M}}(y_i|x) \tag{5.10}$$

4. **Averaged model entropy.** Entropy has also been used to assess the confidence of a model in its own prediction e.g., (Malinin and Gales, 2021; Schröder et al.,

2022; Baumler et al., 2023). A reasonable line of thought is that lower confidence scores reflect data complexity. To evaluate the difficulty of labeling $x$, we can average the label distribution entropy across multiple models:

$$\mathbb{M}_{\text{avg ent}} = -\frac{1}{m}\sum_{j=1}^{m}\sum_{y_i \in Y} p(y_i|x,\theta_j) \times \log p(y_i|x,\theta_j) \tag{5.11}$$

5. **Conformal prediction set size.** A more elaborate statistical estimator than entropy consists in quantifying the ambiguity necessary for a probabilistic classifier to meet a certain statistical guarantee; an approach known as Conformal Prediction (CP) (Vovk et al., 2005; Angelopoulos and Bates, 2022). In practice, we can also use a probabilistic classifier parametrized with $\theta$ to derive a set of possible labels $\mathcal{C}_\theta(x) \subseteq Y$ for every input $x$ such that the true label $y^*$ is likely to be in $\mathcal{C}_\theta(x)$, with a budget tolerance for failure $1-\alpha$. Formally, we want to construct a set-valued function $\mathcal{C}_\theta$ such that

$$\forall x \qquad \Pr\left(y^* \in \mathcal{C}_\theta(x)\right) \geq 1-\alpha$$

We can then capture the ambiguity inherent to a prediction by considering the size of the prediction set, $|\mathcal{C}_\theta(x)|$: a larger CP set size ought to reflect a greater uncertainty as to what the true label is. To convert a probabilistic classifier $p(Y|X,\theta)$ to such a set-valued classifier, we rely on a least-ambiguous set-valued classifier method (Sadinle et al., 2019). This consists in identifying the value $t_\theta$ such that, for all calibration datapoints $x'$ with their label $y'$ in a held-out calibration dataset $\mathcal{D}_{\text{cal}}$:

$$\hat{q} = \frac{|\mathcal{D}_{\text{cal}}|+1}{|\mathcal{D}_{\text{cal}}|}(1-\alpha)$$
$$t_\theta = \sup\left\{t | \Pr\left(p\left(y'|x',\theta\right) \geq t\right) \geq \hat{q}\right\}$$

Using $t_\theta$, we can construct the set

$$\mathcal{C}_\theta(x) = \left\{y \mid p\left(y|x,\theta\right) \geq t_\theta\right\}$$

which provides the expected statistical guarantee. Here, we convert CP sets into uncertainty indicators by considering their average size across models:

$$\mathbb{M}_{\text{CP}} = \frac{1}{m}\sum_{i=1}^{m}|\mathcal{C}_{\theta_i}(x)| \tag{5.12}$$

Here, we experiment with three variants, based on different risk tolerances with $\alpha = 0.05$, $\alpha = 0.1$ or $\alpha = 0.2$. While conformal prediction algorithms require labeled calibration sets $\mathcal{D}_{\text{cal}}$, their predictions are made without label information. Hence we consider CP set size indicators to be reference-free, as they can estimate uncertainty for unlabeled datapoints. We use as $\mathcal{D}_{\text{cal}}$ all other datapoints in the test set (i.e., a leave-one-out process).

6. **Model pool failure rate.** Since it is in principle possible for models to broadly agree on a label that human annotators would not have picked, one value worth considering is the proportion of models that fail to produce the reference label $y^*$ we would expect given our annotations. Defining

$$\mathbb{M}^{\text{ref}}_{\text{fail}} = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\left\{\arg\max_{y_j \in Y} p(y_j|x,\theta_i) \neq y^*\right\} \tag{5.13}$$

highlights the disconnect between model predictions and human annotations. A low value for $\mathbb{M}^{\text{ref}}_{\text{fail}}$ implies a strong alignment between the model pool and the human-provided reference label; a high value suggests that many models fail to predict $y^*$.

7. **Early computation termination.** (Baldock et al., 2021) propose to estimate the difficulty of an example through the computational cost of a correct prediction. They first compute the hidden representations $\mathbf{h}_i^1, \dots \mathbf{h}_i^l$ for a specific input $\mathbf{x}_i$ and then assess which of these representations lie in label-specific subspaces using k-Nearest Neighbor (kNN) classifiers, since datapoints that are easier to process ought to be mapped onto unambiguous subspaces earlier.

   This approach assumes there is a meaningful distance metric between the different representations — an assumption that is not easy to meet with sequence-level classification tasks, where inputs can have different matrix shapes. We can however leverage the fact that Transformer layers can be viewed as functions mapping from and unto the same space (Elhage et al., 2021): Earlier work has suggested to interpret hidden representations for a specific layer by directly projecting them onto the label-space, skipping over all subsequent layers (nostalgebraist, 2020; Geva et al., 2022). We therefore replace (Baldock et al., 2021)'s kNN classifiers with the learned classifier head. More formally, if a model parametrized with $\theta_i$ is made of $l_i$ layers of the form $f_{\theta_{i,j}}(\mathbf{X}) = \phi(\mathbf{X}, \theta_{i,j})$ and a projection head $f_{\theta_i,\text{proj}}(\mathbf{X}) = \arg\max \psi(\mathbf{X}, \theta_{i,l_i+1})$, let us denote all early predictions from layer $j$ onward as

$$\hat{Y}_{ij} = \left\{ f_{\theta_i,\text{proj}} \circ f_{\theta_i,k} \circ \cdots \circ f_{\theta_i,1}(\mathbf{X}) \;\middle|\; j \leq k \leq l_i \right\}$$

   which allows us to retrieve the first layer $k$ such that all predicts from layer $k$ to layer $l$ are correct, according to a reference label $y^*$:

$$\text{s}^{\text{ref}}_{1^{\text{st}}\ \text{layer}}(\theta_i) = \begin{cases} 1 & \text{if } p(y|x,\theta_i) \neq y^* \\ \dfrac{\min_j \{j \mid \hat{Y}_{ij} = \{y^*\}\}}{l+1} & \text{otherwise} \end{cases}$$

$$\mathbb{M}^{\text{ref}}_{1^{\text{st}}\ \text{layer}} = \frac{1}{m}\sum_{i=1}^{m} \text{s}^{\text{ref}}_{1^{\text{st}}\ \text{layer}}(\theta_i) \tag{5.14}$$

   We average across our pool of models so that the indicator is not too sensitive to one specific model's idiosyncratic behavior. This also leads us to normalizing according to the number of layers so that we maintain consistent ranges across models with different layer counts. We also make the practical choice of setting examples that models do not label correctly to the higher end of the scale.

8. **Early training termination.** One can also consider that easier items require less training (Swayamdipta et al., 2020). If for a given model $\theta_i$ we have access to different checkpoints across training $\theta_i^1, \ldots, \theta_i^p$, we can simply assess when the model starts making reliable predictions. Consider the set of predictions from all future checkpoints:

$$F_{ij} = \left\{ \arg\max_y p(y|x, \theta_i^j), \ \ldots, \ \arg\max_y p(y|x, \theta_i^p) \right\}$$

which we use to define:

$$\mathrm{s}^{\text{ref}}_{1^{\text{st}}\ \text{ckpt}}(\theta_i^p) = \begin{cases} 1 & \text{if } p(y|x, \theta_i^p) \neq y^* \\ \frac{\min_j \{j \mid F_{ij} = \{y^*\}\}}{p+1} & \text{otherwise} \end{cases}$$

$$\mathbb{M}^{\text{ref}}_{1^{\text{st}}\ \text{ckpt}} = \frac{1}{m} \sum_{i=1}^{m} \mathrm{s}^{\text{ref}}_{1^{\text{st}}\ \text{ckpt}}(\theta_i^p) \tag{5.15}$$

Here again, we normalize according to the number of checkpoints, average across all models, and manually penalize models that do not ultimately learn to produce the target reference.

9. **Failure rate through training.** We can also assume that easier items are likely to be attributed the expected reference at any stage of training, whereas more complex observations will only be labeled properly during the later stages. We can therefore quantify the proportion of checkpoints where the model failed to produce the expected label $y^*$:

$$\mathbb{M}^{\text{ref}}_{\text{avg}\ \text{ckpt}} = \frac{1}{mp} \sum_{i=1}^{m} \sum_{j=1}^{p} \mathbb{1} \left\{ p(y|x, \theta_i^j) \neq y^* \right\} \tag{5.16}$$

Again, we average across a pool of models to mitigate idiosyncrasies.

10. **Probability mass through training.** One problem with the approach in eq. (5.16) is that it does not distinguish between cases where the classifier correctly predicts $y^*$ and assigns no weight to any other options from cases where the probability assigned to $y^*$ is only within a small margin from that of an incorrect class. (Swayamdipta et al., 2020) propose to consider the probability mass assigned by the classifier across training,[19] or formally:

$$\mathbb{M}^{\text{ref}}_{\text{avg}\ \text{ckpt}\ p} = 1 - \frac{1}{mp} \sum_{i=1}^{m} \sum_{j=1}^{p} p(y^*|x, \theta_i^j) \tag{5.17}$$

Eq. (5.17) is minimized when the gold label $y^*$ is assigned a probability of 1 throughout training.

# Methodology

**Datasets.** We first study classifiers trained on SNLI (Bowman et al., 2015) or MNLI (Williams et al., 2018a) and evaluated on ChaosNLI (Nie et al., 2020b). NLI as a

[19] This indicator corresponds to what (Swayamdipta et al., 2020) call "confidence."

| Dataset | Variant | Train | Val | Test |
|---|---|---|---|---|
| **SNLI** | All labeled (<1B) | 549 367 | 9842 | (*unused*) |
| | 5-splits (1B) | 109 873 | 9842 | (*unused*) |
| **MNLI** | All labeled (<1B) | 392 702 | 9815 | (*unused*) |
| | 5-splits (1B) | 78 540 | 9815 | (*unused*) |
| **ChaosNLI** | SNLI split | — | — | 1000 |
| | MNLI split | — | — | 1000 |

Table 5.17: Dataset statistics.

task is a ternary classification problem, which involves classifying pairs of sentences depending on whether the second sentence contradicts the first; whether the first entails the second; or whether they are neutral with respect to one another, i.e., the second sentence neither derives from nor contradicts the first. We list the statistics as to the number of instances in these datasets in Table 5.17.

**Experimental Setup.** We might expect the family of models we consider to define our indicators to weigh on results. In particular, the homogeneity of the pool of models considered — in terms of pretraining and fine-tuning data, algorithmic and architectural designs, or parameter counts — is a factor of interest.

1. **Heterogeneous training, similar parameter counts (1B group).** One may expect that data complexity indicators should be established by considering a large swath of models trained in conditions as varied as possible — i.e., using different training data and algorithms. To this end, we consider 5 different LLMs in the 1B parameter range; Open Language Model (OLMo) (Groeneveld et al., 2024), Pythia (Biderman et al., 2023), Llama 3.2 (Grattafiori et al., 2024), Falcon (Almazrouei et al., 2023), and BLOOM (Scao et al., 2023). So as to further maximize the difference across the different models we consider, we partition the NLI training set (either SNLI or MNLI) into five equally sized subsets $s_1, \ldots, s_5$ and train one model for each pair of LLM and NLI subset, or 25 classifiers on SNLI and MNLI each.

2. **Homogeneous training data, different parameter counts (<1B group).** Conversely, we might expect that the model pool should be established with a fixed training data — on the one hand, this corresponds to an assumption frequently made when measuring aleatoric uncertainty in the Bayesian literature; on the other hand, we might expect that difficulty should be intimately linked to the data a model has been exposed to. To that tend, we consider a family of smaller BERT-type models (Turc et al., 2019) so as to verify how the indicators in behave with respect to a family of different models trained homogeneously on the same data and under the same conditions, varying in terms of architecture designs and parameter counts. We fine-tune all the 24 (Turc et al., 2019)'s BERT models on each of the NLI training sets in their entirety.

# Results & Discussion

**Human-based and model-based indicators do not agree with each other.** A straightforward first approach consists in computing how the different indicators correlate with one another — in particular, we start by focusing on comparing human-based indicators to model-based indicators.

| | <1B pool | | 1B pool | |
|---|---|---|---|---|
| | $\mathbb{H}_{\text{ent}}$ | $\mathbb{H}_{\text{dis}}$ | $\mathbb{H}_{\text{ent}}$ | $\mathbb{H}_{\text{dis}}$ |
| $\mathbb{M}_{\text{dis}}$ | 0.2440 | 0.2179 | 0.1947 | 0.1772 |
| $\mathbb{M}_{\text{ent}}$ | 0.2784 | 0.2433 | 0.2183 | 0.1970 |
| $\mathbb{M}_{\text{avg ent}}$ | 0.3901 | 0.3490 | 0.2811 | 0.2398 |
| $\mathbb{M}_{\text{CP }\alpha=0.05}$ | 0.3737 | 0.3186 | 0.2767 | 0.2315 |
| $\mathbb{M}_{\text{CP }\alpha=0.1}$ | 0.3763 | 0.3379 | 0.2819 | 0.2393 |
| $\mathbb{M}_{\text{CP }\alpha=0.2}$ | 0.3248 | 0.3064 | 0.2482 | 0.2157 |
| $\mathbb{M}^{\text{ref}}_{\text{fail}}$ | 0.3990 | 0.3959 | 0.3497 | 0.3330 |
| $\mathbb{M}^{\text{ref}}_{\text{1st layer}}$ | 0.3719 | 0.3796 | 0.3624 | 0.3387 |
| $\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$ | 0.4357 | 0.4244 | 0.3682 | 0.3443 |
| $\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$ | 0.3969 | 0.3904 | 0.3477 | 0.3274 |
| $\mathbb{M}^{\text{ref}}_{\text{avg ckpt }p}$ | 0.4386 | 0.4241 | 0.3670 | 0.3428 |

(a) SNLI.

| | <1B pool | | 1B pool | |
|---|---|---|---|---|
| | $\mathbb{H}_{\text{ent}}$ | $\mathbb{H}_{\text{dis}}$ | $\mathbb{H}_{\text{ent}}$ | $\mathbb{H}_{\text{dis}}$ |
| $\mathbb{M}_{\text{dis}}$ | −0.0022 | −0.0045 | 0.1419 | 0.1074 |
| $\mathbb{M}_{\text{ent}}$ | 0.0023 | −0.0011 | 0.1587 | 0.1201 |
| $\mathbb{M}_{\text{avg ent}}$ | −0.0077 | −0.0158 | 0.1329 | 0.1095 |
| $\mathbb{M}_{\text{CP }\alpha=0.05}$ | 0.0101 | −0.0085 | 0.0788 | 0.0425 |
| $\mathbb{M}_{\text{CP }\alpha=0.1}$ | −0.0073 | −0.0173 | 0.1798 | 0.1164 |
| $\mathbb{M}_{\text{CP }\alpha=0.2}$ | −0.0184 | −0.0231 | 0.1581 | 0.0936 |
| $\mathbb{M}^{\text{ref}}_{\text{fail}}$ | 0.1174 | 0.1508 | 0.1726 | 0.2246 |
| $\mathbb{M}^{\text{ref}}_{\text{1st layer}}$ | 0.0682 | 0.0966 | 0.2040 | 0.2514 |
| $\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$ | 0.1132 | 0.1479 | 0.1829 | 0.2307 |
| $\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$ | 0.1168 | 0.1498 | 0.1764 | 0.2261 |
| $\mathbb{M}^{\text{ref}}_{\text{avg ckpt }p}$ | 0.1094 | 0.1434 | 0.1813 | 0.2307 |

(b) MNLI.

Table 5.18: Spearman correlation between human-based and model-based indicators.

The corresponding Spearman correlation values are shown in Tables 5.18a and 5.18b. Reference-free indicators defined without factoring in the majority label among human annotators (eqs. (5.9) to (5.12)) almost systematically yield lower correlations than reference-dependent indicators (eqs. (5.13) to (5.17)). Yet, while we observe positive and significant trends throughout, the correlation itself is somewhat low. For a sense of scale, if we are to focus on SNLI for which we observe the highest correlations, two human-based indicators or two reference-dependent indicators, tend to yield correlation scores of $\rho \geq 0.9$. When comparing two reference-free indicators, we can observe two sub-groups: namely, $\mathbb{M}_{\text{avg ent}}$ and the CP set size indicators yield correlations of $\rho \geq 0.9$),[20] whereas $\mathbb{M}_{\text{dis}}$ and $\mathbb{M}_{\text{ent}}$ yield a correlation of $\rho \approx 0.95$, and comparisons across these two sub-groups are in the range $0.64 < \rho < 0.88$. The observation also holds on MNLI: We observe a correlation of $\rho \approx 0.90$ for $\mathbb{H}_{\text{dis}}$ and $\mathbb{H}_{\text{ent}}$, correlations systematically greater than $\rho \geq 0.9$ between any two reference-dependent indicators, and correlations between $0.46 \leq \rho \leq 0.96$ for reference-free indicators (with again $\mathbb{M}_{\text{dis}}$ and $\mathbb{M}_{\text{ent}}$ forming a subgroup distinct from $\mathbb{M}_{\text{CP}}$ and $\mathbb{M}_{\text{avg ent}}$). In sum, all three groups of indicators portray different pictures, echoing findings from prior works (esp. (Pavlick and Kwiatkowski, 2019)): The difficulty associated to the samples is *not* the same for the humans and models, regardless of the pool considered.[21] The behavior of model-based indicators also appears contingent on the exact setup. For instance, observations derived from our 1B model pool on MNLI would suggest $\mathbb{M}_{\text{CP }\alpha=0.1}$ to be quite in line with human label variation assessments — whereas the corresponding

[20] Except $\mathbb{M}_{\text{CP }\alpha=0.05}$ and $\mathbb{M}_{\text{CP }\alpha=0.2}$, where $\rho \approx 0.80$.

[21] We can stress this relationship is non-linear, see §5.3.2.

coefficient in the <1B pool on MNLI is about 0. In the same vein, the choice of $\alpha$ for CP has different effects on SNLI and MNLI insofar the 1B pool is concerned: Whereas $\mathbb{M}_{\text{CP }\alpha=0.2}$ yields higher results than $\mathbb{M}_{\text{CP }\alpha=0.05}$ on MNLI, the opposite is true for SNLI classifiers.

**Reference-free indicators conflate model success and model failure.** We can also remark that reference-free and reference-dependent indicators do not agree either. This is evident, for instance, by looking at Figure 5.21, which exemplifies one such comparison. We can see that the joint distribution of the indicators forms an inverted U-shape distribution, i.e., the reference-free indicator rates as equally good items that the reference-dependent does discriminates. Generally speaking, reference-free indicators tend to assign similar scores to datapoints rated as either maximal or minimal by reference-dependent indicators: In fact, if we partition datapoints according to whether a majority of the models fail to predict the annotator majority label (corresponding to the orange and blue hues in Figure 5.21), we can observe systematic *anti*-correlations when the models do tend to fail.[22] One major factor at play here is that models fail more often on samples with a high human dissensus. This can be shown with Mann-Whitney U tests. On SNLI, for the 1B models, we observe a $p$-value of $p < 10^{-27}$ and a common language effect size $f = 66.7\%$; as for the <1B models, we have $p < 10^{-42}$, $f = 72.2\%$. On MNLI, the 1B model pool yields $p < 10^{-14}$ and $f = 61.3\%$, whereas the <1B pool yields $p < 10^{-7}$ and $f = 58.1\%$.[23]

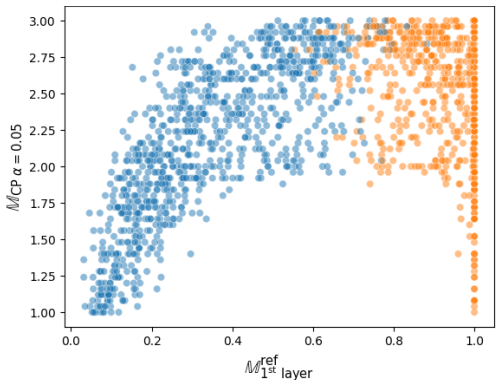



Figure 5.21: Illustration of the joint-distribution reference-free and a reference-dependent indicator.

**Non-linear relationship of human-based indicators and model-based indicators.** To get a better grasp on the magnitude of the difference highlighted in Tables 5.18a and 5.18b, we can turn to residual analyses. We fit a linear regression, attempting to predict one indicator from another, and measure the proportion of variance that this linear model can explain using a coefficient of determination $R^2$. Corresponding values are shown in Table 5.19, with Table 5.19a focusing on the <1B group and Table 5.19b the 1B group. In short, $R^2$ are never above 20%, and often below 10% for reference-free metrics, suggesting that at least 80% of the behavior of our model-based indicators cannot be accounted for with human-based indicators alone. In this case as well, we can observe a difference between reference-free and reference-dependent indicators: As one would expect, reference-dependent indicators yield quantitatively higher $R^2$ scores, suggesting they are (marginally) more in line with human indicators.

[22] See in Tables 5.21 and 5.22 in §5.3.2 for detailed results.

[23] Conversely, we can still identify a small subset of datapoints with low human dissensus but high model failure rates.

| | $\mathbb{H}_{\text{dis}}$ | $\mathbb{H}_{\text{ent}}$ |
|---|---|---|
| $\mathbb{M}_{\text{dis}}$ | 0.0475 | 0.0595 |
| $\mathbb{M}_{\text{ent}}$ | 0.0592 | 0.0775 |
| $\mathbb{M}_{\text{avg ent}}$ | 0.1218 | 0.1521 |
| $\mathbb{M}_{\text{CP } \alpha=0.05}$ | 0.1015 | 0.1396 |
| $\mathbb{M}_{\text{CP } \alpha=0.1}$ | 0.1142 | 0.1416 |
| $\mathbb{M}_{\text{CP } \alpha=0.2}$ | 0.0939 | 0.1055 |
| $\mathbb{M}^{\text{ref}}_{\text{fail}}$ | 0.1568 | 0.1592 |
| $\mathbb{M}^{\text{ref}}_{\text{1st layer}}$ | 0.1441 | 0.1383 |
| $\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$ | 0.1802 | 0.1898 |
| $\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$ | 0.1524 | 0.1576 |
| $\mathbb{M}^{\text{+ref}}_{\text{avg ckpt } p}$ | 0.1799 | 0.1924 |

(a) <1B models

| | $\mathbb{H}_{\text{dis}}$ | $\mathbb{H}_{\text{ent}}$ |
|---|---|---|
| $\mathbb{M}_{\text{dis}}$ | 0.0314 | 0.0379 |
| $\mathbb{M}_{\text{avg ent}}$ | 0.0575 | 0.0790 |
| $\mathbb{M}_{\text{ent}}$ | 0.0388 | 0.0477 |
| $\mathbb{M}_{\text{CP } \alpha=0.05}$ | 0.0536 | 0.0766 |
| $\mathbb{M}_{\text{CP } \alpha=0.1}$ | 0.0573 | 0.0795 |
| $\mathbb{M}_{\text{CP } \alpha=0.2}$ | 0.0465 | 0.0616 |
| $\mathbb{M}^{\text{ref}}_{\text{fail}}$ | 0.1109 | 0.1223 |
| $\mathbb{M}^{\text{ref}}_{\text{1st layer}}$ | 0.1147 | 0.1313 |
| $\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$ | 0.1186 | 0.1356 |
| $\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$ | 0.1072 | 0.1209 |
| $\mathbb{M}^{\text{+ref}}_{\text{avg ckpt } p}$ | 0.1175 | 0.1347 |

(b) 1B models

Table 5.19: Proportion of explained variance ($R^2$) of linear regressions predicting a model-based indicator from a human-based indicator.

| | $\mathbb{M}^{\text{ref}}_{\text{fail}}$ | $\mathbb{M}^{\text{ref}}_{\text{1st layer}}$ | $\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$ | $\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$ | $\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$ |
|---|---|---|---|---|---|
| $\mathbb{M}_{\text{dis}}$ | 0.6190 | 0.5098 | 0.6053 | 0.6091 | 0.5986 |
| $\mathbb{M}_{\text{ent}}$ | 0.6212 | 0.5133 | 0.6083 | 0.6117 | 0.6035 |
| $\mathbb{M}_{\text{avg ent}}$ | 0.5904 | 0.4973 | 0.6221 | 0.5876 | 0.6428 |
| $\mathbb{M}_{\text{CP } \alpha=0.05}$ | 0.4602 | 0.3762 | 0.4845 | 0.4585 | 0.5206 |
| $\mathbb{M}_{\text{CP } \alpha=0.1}$ | 0.4601 | 0.3748 | 0.4881 | 0.4572 | 0.5044 |
| $\mathbb{M}_{\text{CP } \alpha=0.2}$ | 0.3546 | 0.3170 | 0.3747 | 0.3504 | 0.3623 |

(a) <1B models

| | $\mathbb{M}^{\text{ref}}_{\text{fail}}$ | $\mathbb{M}^{\text{ref}}_{\text{1st layer}}$ | $\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$ | $\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$ | $\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$ |
|---|---|---|---|---|---|
| $\mathbb{M}_{\text{dis}}$ | 0.5154 | 0.4966 | 0.5029 | 0.5035 | 0.5048 |
| $\mathbb{M}_{\text{ent}}$ | 0.5168 | 0.4984 | 0.5047 | 0.5073 | 0.5127 |
| $\mathbb{M}_{\text{avg ent}}$ | 0.5292 | 0.5419 | 0.5468 | 0.5292 | 0.5560 |
| $\mathbb{M}_{\text{CP } \alpha=0.05}$ | 0.4860 | 0.4958 | 0.4972 | 0.4916 | 0.5286 |
| $\mathbb{M}_{\text{CP } \alpha=0.1}$ | 0.5216 | 0.5290 | 0.5353 | 0.5241 | 0.5527 |
| $\mathbb{M}_{\text{CP } \alpha=0.2}$ | 0.5232 | 0.5338 | 0.5437 | 0.5188 | 0.5338 |

(b) 1B models

Table 5.20: Spearman correlation between reference-dependent and reference-free indicators.

It is worth highlighting that out of all the reference-free indicators, conformal prediction set sizes and average model entropy scores tend to be the most in line with human judgments. This echoes our earlier remarks on the reference-free indicators being partitioned in two sub-groups, and suggests that more elaborate statistical estimators may mitigate some of the discrepancy we observe between human-based and model-based indicators.

**Interaction of model-based indicators and model success** A more formal statement of the argument shown in Figure 5.21 is that we observe higher correlations when comparing two model-based indicators than when comparing human-based to model-based indicators, although correlations remain smaller than what we observe when comparing indicators within the same group. This can be seen in Table 5.20 for SNLI, where the values are clearly below what we observe within any subgroup of indicators, but higher than what we summarized in Table 5.18a.

To show that all of our reference-free indicators conflate model-success and failure, we can break down observations depending on whether over 50% of the model pool produces the majority label of the human annotator pools. Recomputing correlations

for each subgroup yields systematic negative correlations when the models tend to fail, and systematic positive correlations when the models tend to succeed according to the majority labels, as summarized in Tables 5.21 and 5.22— informally, correlations form the 'left leg' of the inverted U-shape distributions, and anti-correlation the 'right leg.'

| | $\mathbb{M}^{\text{ref}}_{\text{fail}}$ | $\mathbb{M}^{\text{ref}}_{\text{1st layer}}$ | $\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$ | $\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$ | $\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$ |
|---|---|---|---|---|---|
| $\mathbb{M}_{\text{dis}}$ | −0.8508 | −0.8518 | −0.8461 | −0.8085 | −0.7631 |
| $\mathbb{M}_{\text{ent}}$ | −0.7933 | −0.7930 | −0.7863 | −0.7551 | −0.7007 |
| $\mathbb{M}_{\text{avg ent}}$ | −0.5969 | −0.5617 | −0.5390 | −0.5828 | −0.5941 |
| $\mathbb{M}_{\text{CP } \alpha=0.05}$ | −0.3958 | −0.3876 | −0.3874 | −0.3736 | −0.3552 |
| $\mathbb{M}_{\text{CP } \alpha=0.1}$ | −0.4965 | −0.4640 | −0.4670 | −0.4833 | −0.4824 |
| $\mathbb{M}_{\text{CP } \alpha=0.2}$ | −0.4392 | −0.4014 | −0.3974 | −0.4297 | −0.4403 |

(a) <1B models, datapoints where most models fail

| | $\mathbb{M}^{\text{ref}}_{\text{fail}}$ | $\mathbb{M}^{\text{ref}}_{\text{1st layer}}$ | $\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$ | $\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$ | $\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$ |
|---|---|---|---|---|---|
| $\mathbb{M}_{\text{dis}}$ | 0.9507 | 0.6920 | 0.9187 | 0.9213 | 0.8859 |
| $\mathbb{M}_{\text{ent}}$ | 0.9489 | 0.6912 | 0.9177 | 0.9201 | 0.8873 |
| $\mathbb{M}_{\text{avg ent}}$ | 0.8202 | 0.5998 | 0.8760 | 0.8123 | 0.9371 |
| $\mathbb{M}_{\text{CP } \alpha=0.05}$ | 0.5833 | 0.4026 | 0.6342 | 0.5763 | 0.7053 |
| $\mathbb{M}_{\text{CP } \alpha=0.1}$ | 0.6237 | 0.4339 | 0.6775 | 0.6156 | 0.7166 |
| $\mathbb{M}_{\text{CP } \alpha=0.2}$ | 0.4985 | 0.4016 | 0.5317 | 0.4866 | 0.5165 |

(b) <1B models, datapoints where most models succeed

| | $\mathbb{M}^{\text{ref}}_{\text{fail}}$ | $\mathbb{M}^{\text{ref}}_{\text{1st layer}}$ | $\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$ | $\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$ | $\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$ |
|---|---|---|---|---|---|
| $\mathbb{M}_{\text{dis}}$ | −0.7761 | −0.7593 | −0.7737 | −0.7136 | −0.6838 |
| $\mathbb{M}_{\text{ent}}$ | −0.7131 | −0.7075 | −0.7174 | −0.6539 | −0.6140 |
| $\mathbb{M}_{\text{avg ent}}$ | −0.5615 | −0.5264 | −0.5283 | −0.5303 | −0.5111 |
| $\mathbb{M}_{\text{CP } \alpha=0.05}$ | −0.3670 | −0.3633 | −0.3515 | −0.3389 | −0.3037 |
| $\mathbb{M}_{\text{CP } \alpha=0.1}$ | −0.4761 | −0.4565 | −0.4575 | −0.4453 | −0.4182 |
| $\mathbb{M}_{\text{CP } \alpha=0.2}$ | −0.6427 | −0.5836 | −0.5967 | −0.6156 | −0.6116 |

(c) 1B models, datapoints where most models fail

| | $\mathbb{M}^{\text{ref}}_{\text{fail}}$ | $\mathbb{M}^{\text{ref}}_{\text{1st layer}}$ | $\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$ | $\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$ | $\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$ |
|---|---|---|---|---|---|
| $\mathbb{M}_{\text{dis}}$ | 0.9536 | 0.8928 | 0.9159 | 0.9003 | 0.8934 |
| $\mathbb{M}_{\text{ent}}$ | 0.9464 | 0.8891 | 0.9107 | 0.8980 | 0.8962 |
| $\mathbb{M}_{\text{avg ent}}$ | 0.8803 | 0.8955 | 0.9116 | 0.8694 | 0.9315 |
| $\mathbb{M}_{\text{CP } \alpha=0.05}$ | 0.7748 | 0.7939 | 0.7971 | 0.7759 | 0.8546 |
| $\mathbb{M}_{\text{CP } \alpha=0.1}$ | 0.8546 | 0.8601 | 0.8816 | 0.8491 | 0.9103 |
| $\mathbb{M}_{\text{CP } \alpha=0.2}$ | 0.8996 | 0.8982 | 0.9320 | 0.8799 | 0.9166 |

(d) 1B models, datapoints where most models succeed

Table 5.21: Spearman correlation on SNLI data between reference-dependent and reference-free indicators of data complexity, broken down by average model success or failure.

| | $\mathbb{M}^{\text{ref}}_{\text{fail}}$ | $\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$ | $\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$ | $\mathbb{M}^{\text{ref}}_{\text{1st layer}}$ | $\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$ |
|---|---|---|---|---|---|
| $\mathbb{M}_{\text{dis}}$ | −0.7676 | −0.7625 | −0.7812 | −0.6911 | −0.7380 |
| $\mathbb{M}_{\text{ent}}$ | −0.7322 | −0.7301 | −0.7469 | −0.6487 | −0.7034 |
| $\mathbb{M}_{\text{avg ent}}$ | −0.5241 | −0.5130 | −0.5148 | −0.5189 | −0.5044 |
| $\mathbb{M}_{\text{CP } \alpha=0.05}$ | −0.3463 | −0.3307 | −0.3484 | −0.3289 | −0.3320 |
| $\mathbb{M}_{\text{CP } \alpha=0.1}$ | −0.4079 | −0.3930 | −0.4083 | −0.3964 | −0.3930 |
| $\mathbb{M}_{\text{CP } \alpha=0.2}$ | −0.5070 | −0.4949 | −0.5013 | −0.5041 | −0.4915 |

(a) <1B models, datapoints where most models fail

| | $\mathbb{M}^{\text{ref}}_{\text{fail}}$ | $\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$ | $\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$ | $\mathbb{M}^{\text{ref}}_{\text{1st layer}}$ | $\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$ |
|---|---|---|---|---|---|
| $\mathbb{M}_{\text{dis}}$ | 0.9747 | 0.8189 | 0.9593 | 0.9276 | 0.9345 |
| $\mathbb{M}_{\text{ent}}$ | 0.9623 | 0.8053 | 0.9484 | 0.9294 | 0.9281 |
| $\mathbb{M}_{\text{avg ent}}$ | 0.7828 | 0.7236 | 0.8000 | 0.9118 | 0.7530 |
| $\mathbb{M}_{\text{CP } \alpha=0.05}$ | 0.5770 | 0.5819 | 0.5857 | 0.7120 | 0.5664 |
| $\mathbb{M}_{\text{CP } \alpha=0.1}$ | 0.6581 | 0.6282 | 0.6680 | 0.7971 | 0.6466 |
| $\mathbb{M}_{\text{CP } \alpha=0.2}$ | 0.7559 | 0.6753 | 0.7693 | 0.8888 | 0.7386 |

(b) <1B models, datapoints where most models succeed

| | $\mathbb{M}^{\text{ref}}_{\text{fail}}$ | $\mathbb{M}^{\text{ref}}_{\text{1st layer}}$ | $\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$ | $\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$ | $\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$ |
|---|---|---|---|---|---|
| $\mathbb{M}_{\text{dis}}$ | −0.8044 | −0.7910 | −0.8041 | −0.7773 | −0.7784 |
| $\mathbb{M}_{\text{ent}}$ | −0.7594 | −0.7457 | −0.7640 | −0.7346 | −0.7320 |
| $\mathbb{M}_{\text{avg ent}}$ | −0.5422 | −0.4933 | −0.4884 | −0.5259 | −0.5365 |
| $\mathbb{M}_{\text{CP } \alpha=0.05}$ | −0.1415 | −0.1186 | −0.1614 | −0.1246 | −0.1106 |
| $\mathbb{M}_{\text{CP } \alpha=0.1}$ | −0.3913 | −0.3610 | −0.4014 | −0.3724 | −0.3706 |
| $\mathbb{M}_{\text{CP } \alpha=0.2}$ | −0.5293 | −0.4843 | −0.5179 | −0.5155 | −0.5182 |

(c) 1B models, datapoints where most models fail

| | $\mathbb{M}^{\text{ref}}_{\text{fail}}$ | $\mathbb{M}^{\text{ref}}_{\text{1st layer}}$ | $\mathbb{M}^{\text{ref}}_{\text{1st ckpt}}$ | $\mathbb{M}^{\text{ref}}_{\text{avg ckpt}}$ | $\mathbb{M}^{\text{ref}}_{\text{avg ckpt } p}$ |
|---|---|---|---|---|---|
| $\mathbb{M}_{\text{dis}}$ | 0.9767 | 0.8951 | 0.9413 | 0.9447 | 0.9459 |
| $\mathbb{M}_{\text{ent}}$ | 0.9635 | 0.8875 | 0.9309 | 0.9337 | 0.9358 |
| $\mathbb{M}_{\text{avg ent}}$ | 0.7414 | 0.7345 | 0.7863 | 0.7339 | 0.7532 |
| $\mathbb{M}_{\text{CP } \alpha=0.05}$ | 0.0547 | 0.0540 | 0.0366 | 0.0432 | 0.0330 |
| $\mathbb{M}_{\text{CP } \alpha=0.1}$ | 0.5058 | 0.5204 | 0.5220 | 0.4995 | 0.5159 |
| $\mathbb{M}_{\text{CP } \alpha=0.2}$ | 0.6734 | 0.6806 | 0.7091 | 0.6661 | 0.6936 |

(d) 1B models, datapoints where most models succeed

Table 5.22: Spearman correlation on MNLI data between reference-dependent and reference-free indicators of data complexity, broken down by average model success or failure.

Another case worth highlighting concerns $\mathbb{M}_{\text{CP } \alpha=0.05}$ on MNLI within the 1B pool:

correlations and anti-correlations are of remarkably lower magnitudes, which suggests that this specific indicator is not in line with any of the reference-dependent indicators we consider. This is in line with our remarks that the exact setup considered always plays a key role. More broadly, these observations largely confirm our claims in the main body of this article: We find that in most cases, the discrepancy between reference-free and reference-dependent indicators is due to the former conflating model success and model failure.

**Factors shaping model dissensus** We can also leverage the different pools of models to assess how their factors of variation might impact data complexity metrics. In particular, our heterogeneous group of 1B models was defined with respect to different PLMs and training subsets, and therefore we can measure whether pretraining conditions are more impactful than supervised fine-tuning data. In practice, we can measure how likely it is that two predictions for a specific datapoint will match, given that they were made by classifiers trained from the same model or by classifiers trained on the same training subset of SNLI. This can be measured using common language effect sizes derived from a Mann-Whitney U test. Doing so suggests a statistically significant effect from both splits and models ($p < 10^{-44}$) with a very small effect size ($f \approx 51.2\%$) on SNLI. On MNLI, we find a somewhat stronger effect ($f = 53.10\%, p < \epsilon$) when considering classifiers derived from the same PLM; as for the training data, it appears to yield the opposite effect, though with a much higher $p$-value ($f = 49.77\%, p < 10^{-3}$); i.e., any two different PLMs trained on the same split tend to disagree more than other pair of models. In short, there is some evidence that classifiers derived from the same PLM tend to make similar predictions.

Our homogeneous <1B pool also allows us to look into whether responses are more likely to differ for two models with a larger difference in number of parameters. To test this, we can measure the likelihood of the parameter count difference being larger when the predictions differ using U tests. Doing so, we can observe a common language effect size of $f = 45.96\%$ on SNLI and $f = 45.63\%$ on MNLI.We can likewise observe a similar effect when focusing on our heterogeneous pool: we find a common language effect size of $f = 48.90\%$ for SNLI and $f = 45.63\%$ for MNLI.In other words, predictions that match tend to come from models with more similar parameter counts.

Our study of how different indicators of data complexity correlate to one another has shown a somewhat perplexing picture worth diving into. As we have established, model-based indicators align poorly with human-based indicators — while we often observe positive correlations, their magnitudes are low. Defining model-based indicators with respect to human majority labels partially narrows the gap between the two, primarily because reference-free indicators often converge on a single label, regardless of its alignment with human preferences or the strength of consensus within the annotator pool. Within the reference-free indicators, we can also tentatively distinguish two subgroups: assessments that rely only on the pool of models considered (eqs. (5.9) and (5.10)) appear to have a distinct profile from those which rely on more complex statistics, such as CP set sizes or entropy (eqs. (5.11) and (5.12)). For CP specifically, it is worth stressing that desiderata in terms of coverage can also entail significant

variability.

Training classifiers to directly predict human label variation does not fully bridge the gap between reference-free and reference-dependent indicators, and only improves correlations with human assessment for indicators that do not summarize an model's distribution to its argmax. Models often overwhelmingly agree on labels that lack humans annotator consensus, and factoring in human preferences in indicators is necessary though not sufficient for bridging the gap between human-based and model-based assessments difficulty. This underscores a critical limitation of the current research landscape: Reference-free approaches such as CP or entropy are at odds with reference-dependent approaches e.g., (Swayamdipta et al., 2020; Baldock et al., 2021), in that the former conflate failures and successes.[24]

Practical engineering recommendations also emerge from our observations. Authors interested in developing automated assessments of data complexity in line with human assessments should favor (i) training models on soft labels, (ii) factoring in the actual probability distribution of the model, and (iii) leveraging the human label distribution, e.g., through the majority label.

In all, the present observations highlight a disconnect in the current literature. If data uncertainty is to be accounted for by factors such as noise, ambiguity or label overlap during data collection — factors that we also expect to weigh in on measurements of linguistic disagreement — then there is a need to reconcile this line of thought with the limited predictability of model-based assessments of data complexity from annotators' preferences.

## Summary

We present a study with 11 indicators and 29 models, which show that human-based assessments of difficulty need not align with model-based assessments (Pavlick and Kwiatkowski, 2019) and that model-based assessments exhibit stark differences according to whether they factor in human preferences. Data complexity and annotator disagreements, as assessed by model-based indicators or annotator label distribution, have clearly distinct behaviors, despite the overlap the literature posits (Lalor et al., 2018). This calls for replication of our study in other settings, other tasks, other languages, etc.: Establishing the prevalence of the confound we identify remains a topic for future work. Lastly, our findings also question practices adopted by the field. If we are to posit a sharp distinction between data complexity as exemplified by (Swayamdipta et al., 2020) or (Baldock et al., 2021), vs. uncertainty as captured by e.g. conformal prediction methods, then we need to explain why said data complexity is more in line with annotator disagreement than CP-based estimates of uncertainty. Likewise, model-based estimates used in active learning e.g., (Schröder et al., 2022; Baumler et al., 2023) do not align with all definitions of uncertainty, especially label uncertainty as assessed through inter-annotator agreements. Such an exercise in terminology is a

[24] A related train of thought that can shed more light on our observations consists in considering which factors shape model decisions. See §5.3.2 for a discussion.

necessary step forward if we are to address challenges such as disentangling sources of uncertainty (Mucsányi et al., 2024) or leveraging uncertainty as a richer training signal (Basile et al., 2021; Palomaki et al., 2018).

**Limitations** We identify two core limitations on our findings.

First, the present study relies on one dataset, namely the ChaosNLI re-annotation by (Nie et al., 2020b) dataset. While this limits the usefulness of our findings, and entails that our results might not carry on to other setups, we believe this choice is practically necessary (in that very few datasets are available with a training split large enough to easily train classifiers). On a practical level, this also means that there are many human-based indicators that we have ignored; e.g., the 'complicated' label of (Jiang and de Marneffe, 2022) or other explicit self-reports of uncertainty from the annotators could yield valuable insight that would contrast with the label distribution–based indicators we consider in eqs. (5.7) and (5.8); conversely, we have not considered metrics defined with respect to the dataset its in entirety e.g., (Ethayarajh et al., 2022). Of course, all studies need to define their scope: In our case, more can always be done to integrate other data uncertainty/difficulty indicators from a wider range of studies, beyond the key ones we study here (viz. (Nie et al., 2020b; Vovk et al., 2005; Baldock et al., 2021; Swayamdipta et al., 2020)).

Second, we rely on pool of models that have not been individually optimized for the task they are tested on. This point bears further discussion: As we identify in Table 5.20, a major driver of the difference between reference-free and reference-dependent indicators is whether or not the classifier correctly identifies the gold label; and it therefore stands to reason that better trained classifiers may exhibit different patterns. There are however three key facts to stress here. First, hyperparameter tuning over a large pool of models (24 BERT variants from (Turc et al., 2019), plus 25 1B PLMs–based classifiers, on three different datasets) is computationally prohibitive and would actively hinder the reproducibility of our experiments, which justifies the practice of limiting hyperparameter searches. Second, our discussion pertains to the general usefulness of the indicators, rather than the fitness of the models — or in other words, it is reasonable to expect of indicators of data complexity that they be robust enough to be deployed with less-than-top-of-the-leaderboard models. Third, going by Table 5.20, the main driver for the limited correlation between the different groups of model-based indicators is their failure, i.e., the main insight is that we would observe higher correlations if the models never failed, which is not a very realistic standard to expect from NLP systems. While we believe this justifies our approach, it is quite plausible that the exact results as reported here would shift towards higher correlations with human assessments should the models reach higher accuracy scores.

## 5.4 Conclusion

This chapter examines the critical challenges that act as a hurdle in the deployment of medical models in real clinical settings in order to understand *the proficiency gap* between medical models and medical experts. We present the two major views of issues, firstly model-centric issues and secondly data-centric issues. More concretely, in model-centric issues we look into the design choice of a medical model appropriate for medical domain and follow it by investigating the interpretation indicator interaction with the conventional approach of performance focus objective functions. And with data-centric issues we explore the current status of medical concept alignment along with how the data difficulty perception differs between medical models and medical experts.

Across the two dimensions examined in this chapter, a unifying theme emerges: the persistent gap between deep learning or AI-based representations and actual (medical) reality. The proficiency gap illustrates how, despite improvements in predictive accuracy via adding external sources of information which is the focus of this thesis, medical models still remain unable to be at par with the nuanced judgment of medical experts in real-world settings. Model-centric issues such as the single best choice of design between uncertainty-awareness or domain specific models and decrease in interpretability underscore the limitations of the convoluted nature of the modeling paradigm in medical contexts where trust and accountability are highly important. Data-centric challenges, particularly the concept alignment and difficulty perception gaps, further reveal how biases in data collection and representation can distort semantic meaning approximation and lead to systematic errors. Taken together, these findings highlight the intricate interplay between model design, data sources, and clinical applicability, suggesting that progress requires addressing not only individual shortcomings but also their interaction.

In summary, *the proficiency gap*, via the model-centric and data-centric issues together underscore the lingering restrictions that still hinders the trustworthy and widespread integration of medical models into clinical workflows. With this chapter we showcased that these challenges are not completely isolated technical problems but interconnected barriers that reflect deeper tensions between modeling capacity and clinical practice. The next chapter brings the thesis to a close by synthesizing the main findings across all studies, drawing overarching conclusions, and outlining future research directions that can address the identified gaps and guide the development of more trustworthy medical AI systems.

# Part IV

# Conclusion

# Chapter 6

# Conclusion & Future Work



This chapter summarizes the research presented in this thesis, highlighting its key objectives, contributions, and broader implications. We begin by revisiting the goals of the work, followed by a concise summary of the core technical and methodological contributions. The chapter then discusses both theoretical and practical implications of the findings, before acknowledging the limitations of the current study. We conclude with a reflective overview and propose promising directions for future research aimed at advancing the alignment of language models with expert-level reasoning for medical domain.

## 6.1 Thesis Objective

The primary objective of this thesis was to investigate the potential and limits of medical language models through a elaborate analysis of challenges the language models face under its commonly utilization paradigm i.e. single-sourced modeling framework. Throughout the thesis, the experimental design were purposefully kept minimal to facilitate some instant extent of interpretability to enable direct evaluation and identification of the limitations.

This thesis aimed to firstly, systematically examine this gap from multiple perspectives: linguistic complexity (token, sentence, document), and temporal aspect and then explores multi-sourced modeling paradigm by exploring the integration of external knowledge with knowledge bases, knowledge graphs and additional modalities. Through targeted case studies, we explored first hand evidences on how these factors affect the robustness, trustworthiness, and clinical applicability of medical language models. Finally, it concludes by providing a set of reflective studies about why there is a persistent gap between medical models and medical experts.

## 6.2 Key Contributions

This thesis presents several key contributions across different dimensions for the interplay of NLP and the medical domain. We describe them by re-visiting the primary research questions addressed in this thesis:

**What are the challenges faced by Single-sourced Medical Models?** In order to systematically look into in depth the interaction between medical (language) models and medical data, primarily focusing on text data, we took inspiration from the hierarchical view of how language is constructed that is starting from word level then moving to sentence and document level and finally also incorporating the temporal aspect of medical data. Therefore, we performed a total of eight case studies covering all the level of language complexity and temporal aspect of medical data.

We identified several failure cases for medical models at token sentence and document level highlighting issues such as non-standard use of language in social media related medical data; use of shorthand abbreviation, rich context, technical jargon, writing style variation and the need of long context in scientific related medical data. Another common issue is the dependence of medical models on large amount of annotated data.

**Are Multi-Sourced Medical Models more effective compared to Single-sourced Medical Models for medical language processing?** Following the identification of several issues during the study of the previous research question, we performed an exploration of the paradigm of multi sourced medical models via *external knowledge integration* using our single sourced medical models as baselines for a subset of the case studies we

performed during investigation of our first research question. For external knowledge integration, we explored the use of external knowledge base via curated knowledge base for improving robustness against non-standard use of language; retrieval augmented generation paradigm for handling better technical jargons; and scientific citation graph for mitigating the requirement of long context for document understanding. Additionally, we explored the utilization of unstructured clinical modality, in particular clinical notes, to extract complementary information about missing values in clinical data. Finally, we explored a novel paradigm of using LLMs as an assistant to overcome clinical writing style variation.

**Can we explain the proficiency gap between Medical Models and Medical Experts?** With what we achieved during the exploration of multi source modeling paradigm of medical models, this chapter intended to reflect on why there remain a persistent gap between the capability of medical models and a medical human expert (or an imaginary "artificial health expert"). To understand the persistent proficiency gap between medical models and human medical experts, we examined two major perspectives for potential causes: model-centric and data-centric factors.

In model centric issues we looked into the question of what would be the best choice of design for a language model when applying it to a the medical domain with respect to uncertainty and awareness and domain specificity. Additionally, we looked into the interaction of human preference and an interpretability indicator in particular simplicity metric that is used to quantify the usefulness of explanations obtained from the deep learning model. Next for data issues we considered how data is differently perceived by models and human experts. We did so by investigating medical concept alignment and difficulty perception between models and human experts.

This reflective study points to broader methodological and epistemological implications for use of language models (or more generally AI) for medical research. Methodologically, they suggest that current evaluation practices, which often privilege benchmark performance, fail to capture the dimensions of reliability, uncertainty awareness, and interpretability that are central to clinical adoption. Epistemologically, the study reaffirms that existing medical models, unlike human experts, do not embody grounded medical understanding but instead still operate as pattern recognition systems constrained by their training data. These insights call for a more comprehensive framework for evaluating medical AI models, one that integrates technical, clinical, and epistemic considerations rather than treating them as separate domains.

## 6.3 Theoretical & Practical Implications

From a theoretical standpoint, this thesis underscores the need for a more holistic understanding of model performance, one that accounts for not just performance metrics but also alignment with human reasoning, uncertainty calibration, and adaptability to complex medical narratives. This was already recommended from general machine learning paradigm by GoodHart's Law:

*"When a measure becomes a target, if it is effectively optimized, then the thing it is designed to measure will grow worse."*

And, inline with it, our findings argue for a paradigm shift from monolithic evaluation to layered diagnostic frameworks that dissect medical models' strengths and weaknesses, and more generally for domain specific models.

Practically, this research contributes to the development of safer and more robust medical models. Insights into the integration of external knowledge, multimodal data, and LLM assisted usage can inform the design of hybrid architectures that provide both performance and interpretable outputs. Moreover, the emphasis on data-centric and model-centric diagnostics provides a foundation for iterative model refinement, domain-specific benchmarking, and human-in-the-loop deployment strategies.

## 6.4 Limitations

While comprehensive in scope, this thesis is not without limitations. First, the case studies were conducted using relatively simple experimental designs and largely excluded large language models. This decision was influenced both by the timing of the thesis and a deliberate effort to maintain a critical perspective on the rapidly evolving research landscape in the field. Thus, it may not fully cover more recent medical large language models and medical foundational and reasoning models. Additionally, the studies utilized publicly available and often small-sized medical dataset with languages including English, French, Spanish which may not fully capture the diversity of real-world clinical language.

Second, although we explored multiple dimensions of model–expert misalignment, quantitative comparisons with human annotations were limited due to lack of domain experts.

Third, the integration of external knowledge and additional modalities was evaluated in controlled settings, and real-world deployment would involve further challenges such as data synchronization, privacy constraints, system scalability and distributional shift.

## 6.5 Closing Reflection

This thesis began with a simple yet fundamental question: *Why do medical (language) models still struggle to match the reasoning abilities of medical experts?* The exploration that followed revealed that this is not merely a matter of training on more data or building larger models. It is a multifaceted challenge involving linguistic complexity, temporal nature, and systemic biases both in models and in data.

In addressing these issues, this thesis advocates for a move beyond traditional evaluation pipelines, toward frameworks that account for uncertainty, interpretability, modularity, and domain alignment. While language models have immense potential to

augment medical workflows, their trustworthiness must be gained not only through accuracy, but through transparency, and compatibility with the values of medical practice.

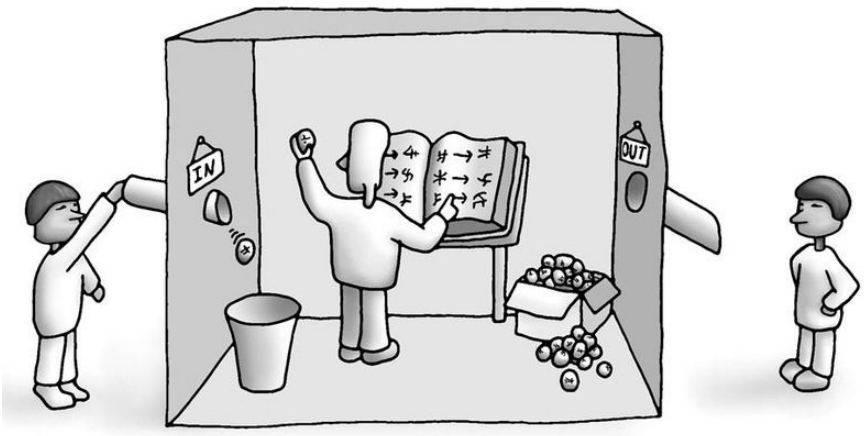



Figure 6.1: The Chinese Room Experiment

In many ways, today's medical language models resemble the system described in Searle's Chinese Room experiment (fig. 6.1): they can manipulate symbols in ways that appear intelligent, even fluent, yet they lack true understanding of the meaning behind those symbols. They may produce clinically plausible responses without grasping medical intent, risk, or nuance. The work presented here underscores that bridging this gap is not just a technical goal but also a conceptual one. It provides a foundation for future research that aims not only to improve model performance, but to deepen our understanding of what proficiency means in medical NLP and how we might move from simulating expertise to cultivating systems that align meaningfully with human judgment and medical reasoning.

## 6.6 Future Direction

Looking ahead, the future of medical NLP is especially inspiring given the success of large language models (LLMs) in general domains. Their capacity to process vast amounts of clinical information and provide targeted, contextually relevant responses could enhance decision support systems, improve patient-doctor communication, accelerate biomedical research, and expand access in low-resource settings. However, despite early successes, careful attention to accuracy, bias, interpretability, and integration into clinical workflows remains essential.

In clinical practice, the use of large language models (LLMs) demands a high degree of factual accuracy, uncertainty awareness, interpretability, and transparency. The "generalist" language models (LMs), trained on humongous web corpora and have undergone task specific tuning already have shown capability to perform better on a wide range of tasks compared to previous approach of specialized pre-trained language mod-

els (PLMs). And, with this cross-task generalization ability, and growing ecosystem of tools and datasets, there is a strong case for investing in the systematic development of medical LLMs. However, these efforts must move beyond simple domain adaptation or fine-tuning. Instead, they must address fundamental concerns around alignment, reasoning, uncertainty quantification, and interpretability all essential properties for safe deployment in clinical practice.

Trivial domain adaptation of general LLMs seems to rarely help for medical setting (Dada et al., 2025) which is in contrast with previous attempts of domain adaptation of PLMs mainly due to the introduction of hallucinations and the decrease of model stability. Hallucination has emerged as a major source of erroneous behavior in LLMs, often arising from knowledge conflicts caused by parametric representations, memorization, or temporal biases (Bi et al., 2025). Clinical reasoning via approaches such as chain-of-thought reasoning alone is not entirely representative of the model's reasoning behind decision making due to lack of tracing back the output to the corpora. Alignment problems in LLMs due to probability-based evaluation strategies (Lyu et al., 2024) often highlights spurious correlations and marks them less close to a reliable system to deploy for healthcare setting.

Future research in medical language models must address several persistent challenges that further hinder their safe and effective deployment in clinical settings. These challenges stem from and compound the issues discussed above such as hallucinations, misalignment, and inadequate domain adaptation. Specifically, the reliance on spurious correlations due to flawed evaluation frameworks, the generation of hallucinated clinical facts, and the absence of reliable mechanisms for expressing uncertainty in high-stakes scenarios all undermine trust in model outputs. Moreover, insufficient grounding in structured medical knowledge, limited interpretability, and vulnerability to domain drift result in models that may appear fluent but are ultimately unreliable. Overcoming these limitations necessitates a paradigm shift toward developing systems that are not only accurate but also inherently trustworthy, transparent, and resilient to the nuanced demands of clinical reasoning. In addition to the above, a rigorous evaluation framework to monitor LLMs' behavior and robustness to subjectivity in human annotation (e.g., LLMs-as-evaluators (Berger et al., 2025), LM-polygraph (Fadeeva et al., 2023)) can be explored to support the trustworthy deployment of LLMs in medical settings.

# Bibliography


Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., and Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297.

Abdill, R. J. and Blekhman, R. (2019). Tracking the popularity and outcomes of all biorxiv preprints. *Elife*, 8:e45133.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Acosta, J. N., Falcone, G. J., Rajpurkar, P., and Topol, E. J. (2022). Multimodal biomedical ai. *Nature medicine*, 28(9):1773–1784.

Adelani, D. (2021). bert-base-multilingual-cased-ner-hrl.

Agarwal, R., Gupta, D., Horsch, A., and Prasad, D. K. (2023a). Aux-drop: Handling haphazard inputs in online learning using auxiliary dropouts. *Transactions on Machine Learning Research*.

Agarwal, R., Sinha, A., Vishwakarma, A., Coubez, X., Clausel, M., Constant, M., Horsch, A., and Prasad, D. K. (2023b). No imputation needed: A switch approach to irregularly sampled time series. *arXiv preprint arXiv:2309.08698*.

AI@Meta (2024). Llama 3 model card. *arXiv preprint arXiv:2407.21783*.

Aji, A. F., Fatyanosa, T. N., Prasojo, R. E., Arthur, P., Fitriany, S., Qonitah, S., Zulfa, N., Santoso, T., and Data, M. (2022). Paracotta: Synthetic multilingual paraphrase corpora from the most diverse translation sample pair. *arXiv preprint arXiv:2205.04651*.

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., and Vollgraf, R. (2019). FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Alangari, N., El Bachir Menai, M., Mathkour, H., and Almosallam, I. (2023). Exploring evaluation methods for interpretable machine learning: A survey. *Information*, 14(8):469.

Albano, A., Di Maria, C., Sciandra, M., and Plaia, A. (2025). Causal forests for discovering diagnostic language in electronic health records. *Applied Stochastic Models in Business and Industry*, 41(5):e70038.

Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Étienne Goffinet, Hesslow, D., Launay, J., Malartic, Q., Mazzotta, D., Noune, B., Pannier, B., and Penedo, G. (2023). The falcon series of open language models.

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., and McDermott, M. (2019a). Publicly available clinical BERT embeddings. In Rumshisky, A., Roberts, K., Bethard, S., and Naumann, T., editors, *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., and McDermott, M. (2019b). Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Alsentzer, E., Rasmussen, M. J., Fontoura, R., Cull, A. L., Beaulieu-Jones, B., Gray, K. J., Bates, D. W., and Kovacheva, V. P. (2023). Zero-shot interpretable phenotyping of postpartum hemorrhage using large language models. *NPJ digital medicine*, 6(1):212.

Alzoubi, H., Alzubi, R., Ramzan, N., West, D., Al-Hadhrami, T., and Alazab, M. (2019). A review of automatic phenotyping approaches using electronic health records. *Electronics*, 8(11):1235.

Angelopoulos, A. N. and Bates, S. (2022). A gentle introduction to conformal prediction and distribution-free uncertainty quantification.

Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Pineda Arango, S., Kapoor, S., Zschiegner, J., Maddix, D. C., Mahoney, M. W., Torkkola, K., Gordon Wilson, A., Bohlke-Schneider, M., and Wang, Y. (2024). Chronos: Learning the language of time series. *Transactions on Machine Learning Research*.

Aronson, A. R. (2001). Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17.

Aroyo, L. and Welty, C. (2015a). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Aroyo, L. and Welty, C. (2015b). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Asai, A., Wu, Z., Wang, Y., Sil, A., and Hajishirzi, H. (2023). Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Atanasova, P. (2024). A diagnostic study of explainability techniques for text classification. In *Accountable and Explainable Methods for Complex Reasoning over Text*, pages 155–187. Springer.

Austin, E., Lee, J. R., Amtmann, D., Bloch, R., Lawrence, S. O., McCall, D., Munson, S., and Lavallee, D. C. (2020). Use of patient-generated health data across healthcare settings: implications for health systems. *JAMIA open*, 3(1):70–76.

Baan, J., Aziz, W., Plank, B., and Fernandez, R. (2022). Stop measuring calibration when humans disagree. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *EMNLP:2022:main*, pages 1892–1915, Abu Dhabi, United Arab Emirates. acl.

Baan, J., Daheim, N., Ilia, E., Ulmer, D., Li, H.-S., Fernández, R., Plank, B., Sennrich, R., Zerva, C., and Aziz, W. (2023). Uncertainty in natural language generation: From theory to applications.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., Hui, B., Ji, L., Li, M., Lin, J., Lin, R., Liu, D., Liu, G., Lu, C., Lu, K., Ma, J., Men, R., Ren, X., Ren, X., Tan, C., Tan, S., Tu, J., Wang, P., Wang, S., Wang, W., Wu, S., Xu, B., Xu, J., Yang, A., Yang, H., Yang, J., Yang, S., Yao, Y., Yu, B., Yuan, H., Yuan, Z., Zhang, J., Zhang, X., Zhang, Y., Zhang, Z., Zhou, C., Zhou, J., Zhou, X., and Zhu, T. (2023). Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Baker, S. and Korhonen, A. (2017). Initializing neural networks for hierarchical multi-label text classification. In Cohen, K. B., Demner-Fushman, D., Ananiadou, S., and Tsujii, J., editors, *Proceedings of the 16th BioNLP Workshop*, pages 307–315, Vancouver, Canada,. Association for Computational Linguistics.

Baker, S., Korhonen, A., and Pyysalo, S. (2016). Cancer hallmark text classification using convolutional neural networks. In Ananiadou, S., Batista-Navarro, R., Cohen, K. B., Demner-Fushman, D., and Thompson, P., editors, *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016)*, pages 1–9, Osaka, Japan. The COLING 2016 Organizing Committee.

Balde, G., Roy, S., Mondal, M., and Ganguly, N. (2025). Evaluation of LLMs in medical text summarization: The role of vocabulary adaptation in high OOV settings. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22989–23004, Vienna, Austria. Association for Computational Linguistics.

Baldock, R. J. N., Maennel, H., and Neyshabur, B. (2021). Deep learning through the lens of example difficulty. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*.

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Bansler, J., Havn, E., Schmidt, K., Mønsted, T., Petersen, H., and Svendsen, J. (2016). Cooperative epistemic work in medical practice: An analysis of physicians' clinical notes. *Computer Supported Cooperative Work*, 25.

Barandas, M., Famiglini, L., Campagner, A., Folgado, D., Simão, R., Cabitza, F., and Gamboa, H. (2024). Evaluation of uncertainty quantification methods in multi-label classification: A case study with automatic diagnosis of electrocardiogram. *Information Fusion*, 101:101978.

Basile, V., Fell, M., Fornaciari, T., Hovy, D., Paun, S., Plank, B., Poesio, M., and Uma, A. (2021). We need to consider disagreement in evaluation. In Church, K., Liberman, M., and Kordoni, V., editors, *BPPF:2021:1*, pages 15–21, Online. acl.

Baumler, C., Sotnikova, A., and Daumé III, H. (2023). Which examples should be multiply annotated? active learning when annotators may disagree. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *FINDINGS:2023:acl*, pages 10352–10371, Toronto, Canada. acl.

Bazoge, A., Morin, E., Daille, B., and Gourraud, P.-A. (2023). Applying natural language processing to textual data from clinical data warehouses: systematic review. *JMIR medical informatics*, 11:e42477.

Beam, A. L., Kompa, B., Schmaltz, A., Fried, I., Weber, G., Palmer, N., Shi, X., Cai, T., and Kohane, I. S. (2020). Clinical concept embeddings learned from massive sources of multimodal medical data. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, volume 25, page 295.

Bedi, S., Jiang, Y., Chung, P., Koyejo, S., and Shah, N. (2025). Fidelity of medical reasoning in large language models. *JAMA Network Open*, 8(8):e2526021–e2526021.

Begoli, E., Bhattacharya, T., and Kusnezov, D. (2019). The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23.

Beigman, E. and Beigman Klebanov, B. (2009). Learning with annotation noise. In Su, K.-Y., Su, J., Wiebe, J., and Li, H., editors, *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 280–287, Suntec, Singapore. Association for Computational Linguistics.

Bejan, C. A., Xia, F., Vanderwende, L., Wurfel, M. M., and Yetisgen-Yildiz, M. (2012). Pneumonia identification using statistical feature selection. *Journal of the American Medical Informatics Association*, 19(5):817–823.

Beltagy, I., Lo, K., and Cohan, A. (2019). Scibert: Pretrained language model for scientific text. In *EMNLP*.

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166.

Berger, A., Khanna, S., Berghaus, D., and Sifa, R. (2025). Reasoning llms in the medical domain: A literature survey. *arXiv preprint arXiv:2508.19097*.

Bhatia, P., Celikkaya, B., and Khalilia, M. (2018). Joint entity extraction and assertion detection for clinical text. *arXiv preprint arXiv:1812.05270*.

Bhatt, U., Weller, A., and Moura, J. M. (2020). Evaluating and aggregating feature-based model explanations. *arXiv preprint arXiv:2005.00631*.

Bi, B., Liu, S., Wang, Y., Xu, Y., Fang, J., Mei, L., and Cheng, X. (2025). Parameters vs. context: Fine-grained control of knowledge reliance in language models. *arXiv preprint arXiv:2503.15888*.

Bian, J., Topaloglu, U., and Yu, F. (2012). Towards large-scale twitter mining for drug-related adverse events. In *Proceedings of the 2012 international workshop on Smart health and wellbeing*, pages 25–32.

Biderman, S., Schoelkopf, H., Anthony, Q., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., Skowron, A., Sutawika, L., and Van Der Wal, O. (2023). Pythia: a suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Blake, C. (2010). Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. *Journal of Biomedical Informatics*, 43(2):173–189.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR.

Boag, W., Doss, D., Naumann, T., and Szolovits, P. (2018). What's in a note? unpacking predictive value in clinical note representations. *AMIA Summits on Translational Science Proceedings*, 2018:26.

Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Van Den Driessche, G. B., Lespiau, J.-B., Damoc, B., Clark, A., et al. (2022). Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Bostrom, K. and Durrett, G. (2020). Byte pair encoding is suboptimal for language model pretraining. *arXiv preprint arXiv:2004.03720*.

Botsis, T., Hartvigsen, G., Chen, F., and Weng, C. (2010). Secondary use of ehr: data quality issues and informatics opportunities. *Summit on translational bioinformatics*, 2010:1.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In Màrquez, L., Callison-Burch, C., and Su, J., editors, *EMNLP:2015:1*, pages 632–642, Lisbon, Portugal. acl.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Buhnila, I. (2023). *Une méthode automatique de construction de corpus de reformulation*. PhD thesis, Université de Strasbourg.

Buhnila, I. and Todirascu, A. (2023). Évaluation d'un générateur automatique de reformulations médicales. In *18e Conférence en Recherche d'Information et Applications\\16e Rencontres Jeunes Chercheurs en RI\\30e Conférence sur le Traitement Automatique des Langues Naturelles\\25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, pages 80–93. ATALA.

Cabitza, F., Rasoini, R., and Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. *Jama*, 318(6):517–518.

Cadwallader, J., Spry, K., Morea, J., Russ, A., Duke, J., and Weiner, M. (2013). Design of a medication reconciliation application. *Applied clinical informatics*, 4(01):110–125.

Canese, K. and Weis, S. (2013). Pubmed: the bibliographic database. *The NCBI handbook*, 2(1):2013.

Carrino, C. P., Armengol-Estapé, J., Gutiérrez-Fandiño, A., Llop-Palao, J., Pàmies, M., Gonzalez-Agirre, A., and Villegas, M. (2021). Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario.

Castro, D. C., Walker, I., and Glocker, B. (2020). Causality matters in medical imaging. *Nature Communications*, 11(1):3673.

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., and Pérez, J. (2020). Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Chalasani, P., Chen, J., Chowdhury, A. R., Wu, X., and Jha, S. (2020). Concise explanations of neural networks using adversarial training. In *International conference on machine learning*, pages 1383–1391. PMLR.

Chancellor, S., Lin, Z., Goodman, E. L., Zerwas, S., and De Choudhury, M. (2016). Quantifying and predicting mental illness severity in online pro-eating disorder communities. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*, pages 1171–1184.

Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., and Buchanan, B. G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310.

Che, Z., Purushotham, S., Cho, K., Sontag, D. A., and Liu, Y. (2016). Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8.

Chen, Z., Subhash, V., Havasi, M., Pan, W., and Doshi-Velez, F. (2022). What makes a good explanation?: A harmonized view of properties of explanations. *arXiv preprint arXiv:2211.05667*.

Chew, C. and Eysenbach, G. (2010). Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak. *PloS one*, 5(11):e14118.

Chizhikova, M., Collado-Montañez, J., López-Úbeda, P., Díaz-Galiano, M. C., López, L. A. U., and Valdivia, M. T. M. (2022). Sinai at clef 2022: Leveraging biomedical transformers to detect and normalize disease mentions. In *CLEF (Working Notes)*, pages 265–273.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014a). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Moschitti, A., Pang, B., and Daelemans, W., editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014b). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Choi, Y., Chiu, C. Y.-I., and Sontag, D. (2016). Learning low-dimensional representations of medical concepts. *AMIA Summits on Translational Science Proceedings*, 2016:41.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. (2023). Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Cogdill, A. P., Andrews, M. C., and Wargo, J. A. (2017). Hallmarks of response to immune checkpoint blockade. *British journal of cancer*, 117(1):1–7.

Collado-Montañez, J., Martín-Valdivia, M.-T., and Martínez-Cámara, E. (2025). Data augmentation based on large language models for radiological report classification. *Knowledge-Based Systems*, 308:112745.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Correia, R. B., Wood, I. B., Bollen, J., and Rocha, L. M. (2020). Mining social media data for biomedical signals and health-related behavior. *Annual review of biomedical data science*, 3(1):433–458.

Cortés, J., Rugo, H. S., Cescon, D. W., Im, S. A., Yusof, M. M., Gallardo, C., Lipatov, O. N., Barrios, C. H., Peréz-García, J. M., Iwata, H., Masuda, N., Otero, M. T., Gokmen, E., Loi, S., Guo, Z., Zhou, X., Karantza, V., Pan, W., and Schmid, P. (2022). Pembrolizumab plus chemotherapy in advanced triple-negative breast cancer. *The New England journal of medicine*, 387 3:217–226.

Costa, F. and Branco, A. (2013). Temporal relation classification based on temporal reasoning. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*, pages 59–70.

Culotta, A., McCallum, A., and Betz, J. (2006). Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In Moore, R. C., Bilmes, J., Chu-Carroll, J., and Sanderson, M., editors, *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 296–303, New York City, USA. Association for Computational Linguistics.

Dada, A., Koraş, O. A., Bauer, M., Corbeil, J.-P., Contreras, A. B., Seibold, C. M., Smith, K. E., Friedrich, J., and Kleesiek, J. (2025). Does biomedical training lead to better medical performance? In Arviv, O., Clinciu, M., Dhole, K., Dror, R., Gehrmann, S., Habba, E., Itzhak, I., Mille, S., Perlitz, Y., Santus, E., Sedoc, J., Shmueli Scheuer, M., Stanovsky, G., and Tafjord, O., editors, *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 46–59, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Dahl, S., Bøgsted, M., Sagi, T., and Vesteghem, C. (2025). Performance of natural language processing for information extraction from electronic health records within cancer: Systematic review. *JMIR Medical Informatics*, 13:e68707.

Dai, J., Upadhyay, S., Aivodji, U., Bach, S. H., and Lakkaraju, H. (2022). Fairness via explanation quality: Evaluating disparities in the quality of post hoc explanations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 203–214.

Daume III, H. and Marcu, D. (2006). Domain adaptation for statistical classifiers. *Journal of artificial Intelligence research*, 26:101–126.

DeepSeek-AI (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Demner-Fushman, D., Chapman, W. W., and McDonald, C. J. (2009). What can natural language processing do for clinical decision support? *Journal of biomedical informatics*, 42(5):760–772.

Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., Thoma, G. R., and McDonald, C. J. (2015). Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2024). Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Deznabi, I., Iyyer, M., and Fiterau, M. (2021). Predicting in-hospital mortality by combining clinical notes with time-series data. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pages 4026–4031.

Dolan, W., Quirk, C., Brockett, C., and Dolan, B. (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*.

Donnelly, K. et al. (2006). Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Džuganová, B. (2019). Medical language-a unique linguistic phenomenon: Engleski. *Jahr–European Journal of Bioethics*, 10(1):129–145.

Eddine, M. K., Tixier, A. J.-P., and Vazirgiannis, M. (2020). Barthez: a skilled pretrained french sequence-to-sequence model. *arXiv preprint arXiv:2010.12321*.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. (2021). A mathematical framework for transformer circuits.

Ermakova, L., Bordignon, F., Turenne, N., and Noel, M. (2018). Is the abstract a mere teaser? evaluating generosity of article abstracts in the environmental sciences. *Frontiers Res. Metrics Anal.*, 3:16.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., and Dean, J. (2019). A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29.

Ethayarajh, K., Choi, Y., and Swayamdipta, S. (2022). Understanding dataset difficulty with $\mathcal{V}$-usable information. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

Fadeeva, E., Vashurin, R., Tsvigun, A., Vazhentsev, A., Petrakov, S., Fedyanin, K., Vasilev, D., Goncharova, E., Panchenko, A., Panov, M., Baldwin, T., and Shelmanov, A. (2023). LM-polygraph: Uncertainty estimation for language models. In Feng, Y. and Lefever, E., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 446–461, Singapore. Association for Computational Linguistics.

Ferraro, J. P., Daumé III, H., DuVall, S. L., Chapman, W. W., Harkema, H., and Haug, P. J. (2013). Improving performance of natural language processing part-of-speech tagging on clinical narratives through domain adaptation. *Journal of the American Medical Informatics Association*, 20(5):931–939.

Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I. S., and Saria, S. (2021). The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, 385(3):283–286.

Fivez, P., Suster, S., and Daelemans, W. (2021). Conceptual grounding constraints for truly robust biomedical name representations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2440–2450.

Flamholz, Z. N., Crane-Droesch, A., Ungar, L. H., and Weissman, G. E. (2022). Word embeddings trained on published case reports are lightweight, effective for clinical tasks, and free of protected health information. *Journal of Biomedical Informatics*, 125:103971.

Fornaciari, T., Uma, A., Paun, S., Plank, B., Hovy, D., and Poesio, M. (2021). Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *NAACL:2021:main*, pages 2591–2597, Online. acl.

Frantar, E., Ashkboos, S., Hoefler, T., and Alistarh, D. (2022). Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.

Friedman, C. (1997). Towards a comprehensive medical language processing system: methods and issues. In *Proceedings of the AMIA annual fall symposium*, page 595.

Fu, S., Chen, D., He, H., Liu, S., Moon, S., Peterson, K. J., Shen, F., Wang, L., Wang, Y., Wen, A., et al. (2020). Clinical concept extraction: a methodology review. *Journal of biomedical informatics*, 109:103526.

Fu, Y., Ramachandran, G. K., Dobbins, N. J., Park, N., Leu, M., Rosenberg, A. R., Lybarger, K., Xia, F., Uzuner, O., and Yetisgen, M. (2024). Extracting social determinants of health from pediatric patient notes using large language models: Novel corpus and methods. *ArXiv*, abs/2404.00826.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Ganitkevitch, J. and Callison-Burch, C. (2014). The multilingual paraphrase database. In *LREC*, pages 4276–4283. Citeseer.

Gao, Y., Li, R., Croxford, E., Caskey, J., Patterson, B. W., Churpek, M., Miller, T., Dligach, D., and Afshar, M. (2025). Leveraging medical knowledge graphs into large language models for diagnosis prediction: Design and application study. *Jmir Ai*, 4:e58670.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., and Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., and Zhu, X. X. (2023). A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(1):1513–1589.

Geng, J., Cai, F., Wang, Y., Koeppl, H., Nakov, P., and Gurevych, I. (2024). A survey of confidence estimation and calibration in large language models. In Duh, K., Gomez, H., and Bethard, S., editors, *NAACL:2024:long*, pages 6577–6595, Mexico City, Mexico. acl.

Gennari, A., André, F., Barrios, C., Cortes, J., de Azambuja, E., DeMichele, A., Dent, R., Fenlon, D., Gligorov, J., Hurvitz, S., et al. (2021). Esmo clinical practice guideline for the diagnosis, staging and treatment of patients with metastatic breast cancer. *Annals of oncology*, 32(12):1475–1495.

Geva, M., Caciularu, A., Wang, K., and Goldberg, Y. (2022). Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *EMNLP:2022:main*, pages 30–45, Abu Dhabi, United Arab Emirates. acl.

Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y., and Ranganath, R. (2020). A review of challenges and opportunities in machine learning for health. *AMIA Summits on Translational Science Proceedings*, 2020:191.

Ginn, R., Pimpalkhute, P., Nikfarjam, A., Patki, A., O'Connor, K., Sarker, A., Smith, K., and Gonzalez, G. (2014). Mining twitter for adverse drug reaction mentions: a corpus and classification benchmark. In *Proceedings of the fourth workshop on building and evaluating resources for health and biomedical text processing*, pages 1–8.

Gkotsis, G., Velupillai, S., Oellrich, A., Dean, H., Liakata, M., and Dutta, R. (2016). Don't let notes be misunderstood: A negation detection method for assessing risk of suicide in mental health records. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 95–105.

Goldberg, E. (1931). Statistical machine. US Patent 1,838,389.

Gomez, T., Fréour, T., and Mouchère, H. (2022). Comparison of attention models and post-hoc explanation methods for embryo stage identification: a case study. In *International Conference on Pattern Recognition*, pages 216–230. Springer.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Goodman, N. W. and Edwards, M. B. (2014). *Medical writing: a prescription for clarity*. Cambridge University Press.

Grabar, N. and Cardon, R. (2018). Clear-simple corpus for medical french. In *ATA*.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R.,

Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon,

S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. (2024). The llama 3 herd of models.

Groeneveld, D., Beltagy, I., Walsh, E., Bhagia, A., Kinney, R., Tafjord, O., Jha, A., Ivison, H., Magnusson, I., Wang, Y., Arora, S., Atkinson, D., Authur, R., Chandu, K., Cohan, A., Dumas, J., Elazar, Y., Gu, Y., Hessel, J., Khot, T., Merrill, W., Morrison, J., Muennighoff, N., Naik, A., Nam, C., Peters, M., Pyatkin, V., Ravichander, A., Schwenk, D., Shah, S., Smith, W., Strubell, E., Subramani, N., Wortsman, M., Dasigi, P., Lambert, N., Richardson, K., Zettlemoyer, L., Dodge, J., Lo, K., Soldaini, L., Smith, N., and Hajishirzi, H. (2024). OLMo: Accelerating the science of language models. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *ACL:2024:long*, pages 15789–15809, Bangkok, Thailand. acl.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2):199–220.

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., et al. (2024). A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2021a). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2021b). Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Gumiel, Y. B., Silva e Oliveira, L. E., Claveau, V., Grabar, N., Paraiso, E. C., Moro, C., and Carvalho, D. R. (2021). Temporal relation extraction in clinical texts: a systematic review. *ACM Computing Surveys (CSUR)*, 54(7):1–36.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Guo, Y., Korhonen, A., Liakata, M., Silins, I., Högberg, J., and Stenius, U. (2011). A comparison and user-based evaluation of models of textual information structure in the context of cancer risk assessment. *BMC Bioinformatics*, 12:69 – 69.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M. (2020). Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Hachey, B., Alex, B., and Becker, M. (2005). Investigating the effects of selective sampling on the annotation task. In Dagan, I. and Gildea, D., editors, *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 144–151, Ann Arbor, Michigan. Association for Computational Linguistics.

Hamilton, W., Ying, Z., and Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Yao, Y., Zhang, A., Zhang, L., et al. (2021). Pre-trained models: Past, present and future. *Ai Open*, 2:225–250.

Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *cell*, 100(1):57–70.

Harkema, H., Dowling, J. N., Thornblade, T., and Chapman, W. W. (2009). Context: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of biomedical informatics*, 42(5):839–851.

Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G., and Galstyan, A. (2019). Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96.

Hayat, N., Geras, K. J., and Shamout, F. E. (2022). Medfuse: Multi-modal fusion with clinical time-series data and chest x-ray images. In *Machine Learning for Healthcare Conference*, pages 479–503. PMLR.

He, Y., Wang, C., Zhang, S., Li, N., Li, Z., and Zeng, Z. (2022). Kg-mtt-bert: knowledge graph enhanced bert for multi-type medical text classification. *arXiv preprint arXiv:2210.03970*.

Hedström, A., Bommer, P., Wickstrøm, K. K., Samek, W., Lapuschkin, S., and Höhne, M. M.-C. (2023). The meta-evaluation problem in explainable ai: identifying reliable estimators with metaquantus. *arXiv preprint arXiv:2302.07265*.

Hendrycks, D. and Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*.

Hirschberg, J. and Manning, C. D. (2015). Advances in natural language processing. *Science*, 349(6245):261–266.

Ho, T. K. and Basu, M. (2002). Complexity measures of supervised classification problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):289–300.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Holzinger, A., Biemann, C., Pattichis, C. S., and Kell, D. B. (2017). What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*.

Horn, M., Moor, M., Bock, C., Rieck, B., and Borgwardt, K. (2020). Set functions for time series. In *International Conference on Machine Learning*, pages 4353–4363. PMLR.

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Hu, M., Zhang, Z., Zhao, S., Huang, M., and Wu, B. (2023). Uncertainty in natural language processing: Sources, quantification, and applications. *arXiv preprint arXiv:2306.04459*.

Huang, B. (2023). Vigogne: French instruction-following and chat models. https://github.com/bofenghuang/vigogne.

Huang, K., Altosaar, J., and Ranganath, R. (2019). Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Huang, P.-S., He, X., Gao, J., Deng, L., Acero, A., and Heck, L. (2013). Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2333–2338.

Huang, S.-C., Pareek, A., Seyyedi, S., Banerjee, I., and Lungren, M. P. (2020). Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ digital medicine*, 3(1):136.

Hüllermeier, E. and Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3):457–506.

Hurd, M. D., Martorell, P., Delavande, A., Mullen, K. J., and Langa, K. M. (2013). Monetary costs of dementia in the united states. *New England Journal of Medicine*, 368(14):1326–1334.

Hwang, J., Gudumotu, C., and Ahmadnia, B. (2023). Uncertainty quantification of text classification in a multi-label setting for risk-sensitive systems. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 541–547.

Ilievski, I. and Feng, J. (2017). Multimodal learning and reasoning for visual question answering. *Advances in neural information processing systems*, 30.

Ipsen, N. B., Mattei, P.-A., and Frellsen, J. (2020). not-miwae: Deep generative modelling with missing not at random data. In *International Conference on Learning Representations*.

Izacard, G. and Grave, E. (2021). Leveraging passage retrieval with generative models for open domain question answering. In Merlo, P., Tiedemann, J., and Tsarfaty, R., editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Jagannatha, A., Liu, F., Liu, W., and Yu, H. (2019). Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (made 1.0). *Drug safety*, 42(1):99–111.

Jamison, E. and Gurevych, I. (2015). Noise or additional information? leveraging crowdsource annotation item agreement for natural language tasks. In Màrquez, L., Callison-Burch, C., and Su, J., editors, *EMNLP:2015:1*, pages 291–297, Lisbon, Portugal. acl.

Jensen, P. B., Jensen, L. J., and Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405.

Jeong, M., Sohn, J., Sung, M., and Kang, J. (2024). Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *arXiv preprint arXiv:2401.15269*.

Ji, S., Pan, S., Cambria, E., Marttinen, P., and Yu, P. S. (2021). A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems*, 33(2):494–514.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Jiang, N.-J. and de Marneffe, M.-C. (2022). Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., and Szolovits, P. (2021). What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Jin, Q., Dhingra, B., Liu, Z., Cohen, W., and Lu, X. (2019a). PubMedQA: A dataset for biomedical research question answering. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Jin, W., Zhang, C., Szekely, P. A., and Ren, X. (2019b). Recurrent event network for reasoning over temporal knowledge graphs. *ArXiv*, abs/1904.05530.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Johnson, R., Watkinson, A., and Mabe, M. (2018). The stm report: An overview of scientific and scholarly publishing. *STM: International Association of Scientific, Technical and Medical Publishers*.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., and Mikolov, T. (2016). Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Jurafsky, D. and Martin, J. H. (2023). *Speech and Language Processing*. Pearson, 3rd edition. Draft chapters available online: https://web.stanford.edu/~jurafsky/slp3/.

Kanakarajan, K. r., Kundumani, B., and Sankarasubbu, M. (2021). BioELECTRA:pretrained biomedical text encoder using discriminators. In Demner-Fushman, D., Cohen, K. B., Ananiadou, S., and Tsujii, J., editors, *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 143–154, Online. Association for Computational Linguistics.

Kandula, S., Curtis, D., and Zeng-Treitler, Q. (2010). A semantic and syntactic text simplification tool for health content. In *AMIA annual symposium proceedings*, volume 2010, page 366.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Karasarides, M., Cogdill, A. P., Robbins, P. B., Bowden, M., Burton, E. M., Butterfield, L. H., Cesano, A., Hammer, C., Haymaker, C. L., Horak, C. E., et al. (2022). Hallmarks of resistance to immune-checkpoint inhibitors. *Cancer immunology research*, 10(4):372–383.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Kazemi, S. M., Goel, R., Eghbali, S., Ramanan, J., Sahota, J., Thakur, S., Wu, S., Smyth, C., Poupart, P., and Brubaker, M. (2019). Time2vec: Learning a vector representation of time. *arXiv preprint arXiv:1907.05321*.

Kendall, A. and Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5580–5590, Red Hook, NY, USA. Curran Associates Inc.

Khadanga, S., Aggarwal, K., Joty, S., and Srivastava, J. (2019). Using clinical notes with time series data for icu management. *arXiv preprint arXiv:1909.09702*.

Khatir, M. and Reddy, C. K. (2024). Concept formation and alignment in language models: Bridging statistical patterns in latent space to concept taxonomy. *arXiv preprint arXiv:2406.05315*.

Kilicoglu, H., Shin, D., Fiszman, M., Rosemblat, G., and Rindflesch, T. C. (2012). Semmeddb: a pubmed-scale repository of biomedical semantic predications. *Bioinformatics*, 28(23):3158–3160.

Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003). Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182.

Kim, S., Lee, N., Lee, J., Hyun, D., and Park, C. (2023). Heterogeneous graph learning for multi-modal medical data analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 5141–5150.

Kipf, T. and Welling, M. (2017). Semi-supervised classification with graph convolutional networks. *ArXiv*, abs/1609.02907.

Kiureghian, A. D. and Ditlevsen, O. (2009). Aleatory or epistemic? does it matter? *Structural Safety*, 31(2):105–112. Risk Acceptance and Risk Communication.

Kline, A., Wang, H., Li, Y., Dennis, S., Hutch, M., Xu, Z., Wang, F., Cheng, F., and Luo, Y. (2022). Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine*, 5(1):171.

Klug, K., Beckh, K., Antweiler, D., Chakraborty, N., Baldini, G., Laue, K., Hosch, R., Nensa, F., Schuler, M., and Giesselbach, S. (2024). From admission to discharge: a

systematic review of clinical natural language processing along the patient journey. *BMC Medical Informatics and Decision Making*, 24(1):238.

Kocabiyikoglu, A., Portet, F., Gibert, P., Blanchon, H., Babouchkine, J.-M., and Gavazzi, G. (2022). A spoken drug prescription dataset in french for spoken language understanding. In *13th Language Resources and Evaluation Conference (LREC 2022)*.

Köhler, S., Vasilevsky, N. A., Engelstad, M., Foster, E., McMurry, J., Aymé, S., Baynam, G., Bello, S. M., Boerkoel, C. F., Boycott, K. M., et al. (2017). The human phenotype ontology in 2017. *Nucleic acids research*, 45(D1):D865–D876.

Kraljevic, Z., Searle, T., Shek, A., Roguski, L., Noor, K., Bean, D., Mascio, A., Zhu, L., Folarin, A. A., Roberts, A., et al. (2021). Multi-domain clinical natural language processing with medcat: the medical concept annotation toolkit. *Artificial intelligence in medicine*, 117:102083.

Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S. F., Forshee, R., Walderhaug, M., and Botsis, T. (2017). Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *Journal of biomedical informatics*, 73:14–29.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

Kuchibhotla, A. K., Brown, L. D., Buja, A., and Cai, J. (2019). All of linear regression. *arXiv preprint arXiv:1910.06386*.

Kudo, T. and Richardson, J. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Kumar, V., Stubbs, A., Shaw, S., and Uzuner, Ö. (2015). Creation of a new longitudinal corpus of clinical narratives. *Journal of biomedical informatics*, 58:S6–S10.

Kundeti, S. R., Vijayananda, J., Mujjiga, S., and Kalyan, M. (2016). Clinical named entity recognition: Challenges and opportunities. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 1937–1945. IEEE.

Kwon, S., Yao, Z., Jordan, H., Levy, D., Corner, B., and Yu, H. (2022). MedJEx: A medical jargon extraction model with Wiki's hyperlink span and contextualized masked language model score. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11733–11751, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Labrak, Y., Bazoge, A., Dufour, R., Rouvier, M., Morin, E., Daille, B., and Gourraud, P.-A. (2023a). Drbert: A robust pre-trained model in french for biomedical and clinical domains. *bioRxiv*, pages 2023–04.

Labrak, Y., Bazoge, A., Dufour, R., Rouvier, M., Morin, E., Daille, B., and Gourraud, P.-A. (2023b). Frenchmedmcqa: A french multiple-choice question answering dataset for medical domain. *arXiv preprint arXiv:2304.04280*.

Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.-A., Rouvier, M., and Dufour, R. (2024). Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

Labrak, Y., Rouvier, M., and Dufour, R. (2023c). MORFITT : Un corpus multi-labels d'articles scientifiques français dans le domaine biomédical. In Boudin, F., Daille, B., Dufour, R., Khettari, O., Houbre, M., Jourdan, L., and Kooli, N., editors, *18e Conférence en Recherche d'Information et Applications – 16e Rencontres Jeunes Chercheurs en RI – 30e Conférence sur le Traitement Automatique des Langues Naturelles – 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, pages 66–70, Paris, France. ATALA.

Labrak, Y., Rouvier, M., and Dufour, R. (2023d). A zero-shot and few-shot study of instruction-finetuned large language models applied to clinical and biomedical tasks. *arXiv preprint arXiv:2307.12114*.

Lai, T. M., Zhai, C., and Ji, H. (2023). Keblm: knowledge-enhanced biomedical language models. *Journal of Biomedical Informatics*, 143:104392.

Lalor, J. P., Wu, H., Munkhdalai, T., and Yu, H. (2018). Understanding deep learning performance through an examination of test set difficulty: A psychometric case study. *Proc. Conf. Empir. Methods Nat. Lang. Process.*, 2018:4711–4716.

Lange, L., Adel, H., Strötgen, J., and Klakow, D. (2021). Clin-x: pre-trained language models and a study on cross-task transfer for concept extraction in the clinical domain.

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A. C., Liu, Y., Maciejewski, A., Arndt, D., Wilson, M., Neveu, V., et al. (2014). Drugbank 4.0: shedding new light on drug metabolism. *Nucleic acids research*, 42(D1):D1091–D1097.

Leaman, R., Khare, R., and Lu, Z. (2015). Challenges in clinical natural language processing for automated disorder normalization. *Journal of biomedical informatics*, 57:28–37.

Lee, B. W. and Lee, J. (2023). LFTK: Handcrafted features in computational linguistics. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 1–19, Toronto, Canada. Association for Computational Linguistics.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Lewis, P., Ott, M., Du, J., and Stoyanov, V. (2020a). Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art.

In Rumshisky, A., Roberts, K., Bethard, S., and Naumann, T., editors, *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157, Online. Association for Computational Linguistics.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020b). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Lhoest, Q., Villanova del Moral, A., Jernite, Y., Thakur, A., von Platen, P., Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L., Davison, J., Šaško, M., Chhablani, G., Malik, B., Brandeis, S., Le Scao, T., Sanh, V., Xu, C., Patry, N., McMillan-Major, A., Schmid, P., Gugger, S., Delangue, C., Matussière, T., Debut, L., Bekman, S., Cistac, P., Goehringer, T., Mustar, V., Lagunas, F., Rush, A., and Wolf, T. (2021). Datasets: A community library for natural language processing. In Adel, H. and Shi, S., editors, *EMNLP:2021:demo*, pages 175–184, Online and Punta Cana, Dominican Republic. acl.

Li, F., Jin, Y., Liu, W., Rawat, B. P. S., Cai, P., and Yu, H. (2019a). Fine-tuning bidirectional encoder representations from transformers (bert)–based models on large-scale electronic health record notes: an empirical study. *JMIR medical informatics*, 7(3):e14830.

Li, J., Sun, Y., Johnson, R. J., Sciaky, D., Wei, C.-H., Leaman, R., Davis, A. P., Mattingly, C. J., Wiegers, T. C., and Lu, Z. (2016). Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Li, L., Xu, J., Dong, Q., Zheng, C., Sun, X., Kong, L., and Liu, Q. (2023). Can language models understand physical concepts? In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11843–11861, Singapore. Association for Computational Linguistics.

Li, X., Sun, X., Meng, Y., Liang, J., Wu, F., and Li, J. (2019b). Dice loss for data-imbalanced nlp tasks. *arXiv preprint arXiv:1911.02855*.

Li, Y., Rao, S., Solares, J. R. A., Hassaine, A., Ramakrishnan, R., Canoy, D., Zhu, Y., Rahimi, K., and Salimi-Khorshidi, G. (2020). Behrt: transformer for electronic health records. *Scientific reports*, 10(1):7155.

Li, Y., Wehbe, R. M., Ahmad, F. S., Wang, H., and Luo, Y. (2022). Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *arXiv preprint arXiv:2201.11838*.

Li, Z., Li, S., and Yan, X. (2024). Time series as images: Vision transformer for irregularly sampled time series. *Advances in Neural Information Processing Systems*, 36.

Lim, B., Arık, S. Ö., Loeff, N., and Pfister, T. (2021a). Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764.

Lim, D. K., Rashid, N. U., Oliva, J. B., and Ibrahim, J. G. (2021b). Handling non-ignorably missing features in electronic health records data using importance-weighted autoencoders. *arXiv e-prints*, pages arXiv–2101.

Limsopatham, N. and Collier, N. (2016). Normalising medical concepts in social media texts by learning semantic representation. In Erk, K. and Smith, N. A., editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1014–1023, Berlin, Germany. Association for Computational Linguistics.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out, Barcelona, Spain. Association for Computational Linguistics, 2004*, page 74–81.

Lin, J. J. (2009). Is searching full text more effective than searching abstracts? *BMC Bioinformatics*, 10:46 – 46.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Liu, F., Li, Z., Zhou, H., Yin, Q., Yang, J., Tang, X., Luo, C., Zeng, M., Jiang, H., Gao, Y., Nigam, P., Nag, S., Yin, B., Hua, Y., Zhou, X., Rohanian, O., Thakur, A., Clifton, L., and Clifton, D. A. (2024). Large language models in the clinic: A comprehensive benchmark.

Liu, W., Zhou, P., Zhao, Z., Wang, Z., Ju, Q., Deng, H., and Wang, P. (2020). K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 2901–2908.

Liu, Z., Alavi, A., Li, M., and Zhang, X. (2023). Self-supervised contrastive learning for medical time series: A systematic review. *Sensors*, 23(9):4221.

Liu, Z., Zhang, J., Hou, Y., Zhang, X., Li, G., and Xiang, Y. (2022). Machine learning for multimodal electronic health records-based research: Challenges and perspectives. In *China Health Information Processing Conference*, pages 135–155. Springer.

Llanos, L. C., Bouamor, D., Zweigenbaum, P., and Rosset, S. (2016a). Managing linguistic and terminological variation in a medical dialogue system. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation*

*(LREC'16)*, pages 3167–3173, Portorož, Slovenia. European Language Resources Association (ELRA).

Llanos, L. C., Bouamor, D., Zweigenbaum, P., and Rosset, S. (2016b). Managing linguistic and terminological variation in a medical dialogue system. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3167–3173.

Logan, R., Liu, N. F., Peters, M. E., Gardner, M., and Singh, S. (2019). Barack's wife hillary: Using knowledge graphs for fact-aware language modeling. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5962–5971, Florence, Italy. Association for Computational Linguistics.

Loibl, S., André, F., Bachelot, T., Barrios, C., Bergh, J., Burstein, H., Cardoso, M., Carey, L., Dawood, S., Del Mastro, L., et al. (2024). Early breast cancer: Esmo clinical practice guideline for diagnosis, treatment and follow-up. *Annals of Oncology*, 35(2):159–182.

Lorena, A. C., Garcia, L. P. F., Lehmann, J., Souto, M. C. P., and Ho, T. K. (2019). How complex is your classification problem? a survey on measuring classification complexity. *ACM Comput. Surv.*, 52(5).

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Lu, Z. (2011). Pubmed and beyond: a survey of web tools for searching biomedical literature. *Database*, 2011:baq036.

Lundberg, S. M. and Lee, S.-I. (2017a). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Lundberg, S. M. and Lee, S.-I. (2017b). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., and Liu, T.-Y. (2022). BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

Lyu, C., Wu, M., and Aji, A. (2024). Beyond probabilities: Unveiling the misalignment in evaluating large language models. In Li, S., Li, M., Zhang, M. J., Choi, E., Geva, M., Hase, P., and Ji, H., editors, *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 109–131, Bangkok, Thailand. Association for Computational Linguistics.

Ma, C. and Zhang, C. (2021). Identifiable generative models for missing not at random data imputation. *Advances in Neural Information Processing Systems*, 34:27645–27658.

Madsen, A., Reddy, S., and Chandar, S. (2022). Post-hoc interpretability for neural nlp: A survey. *ACM Comput. Surv.*, 55(8).

Mahajan, D., Liang, J. J., and Tsou, C. (2020). Toward understanding clinical context of medication change events in clinical narratives. *CoRR*, abs/2011.08835.

Mahajan, D., Liang, J. J., and Tsou, C.-H. (2022). Toward Understanding Clinical Context of Medication Change Events in Clinical Narratives. *AMIA Annual Symposium Proceedings*, 2021:833–842.

Mahajan, D., Liang, J. J., Tsou, C.-H., and Uzuner, Ö. (2023). Overview of the 2022 n2c2 shared task on contextualized medication event extraction in clinical notes. *Journal of biomedical informatics*, 144:104432.

Malinin, A. and Gales, M. (2021). Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*.

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de La Clergerie, É. V., Seddah, D., and Sagot, B. (2019). Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894*.

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, É. V., Seddah, D., and Sagot, B. (2020). Camembert: a tasty french language mode. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., and Mukherjee, A. (2021). Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

McCoy, R. T., Pavlick, E., and Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*.

Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., and Hurdle, J. F. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of medical informatics*, 17(01):128–144.

Mickus, T., Sinha, A., and Vázquez, R. (2025). Your model is overconfident, and other lies we tell ourselves. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5401–5417, Vienna, Austria. Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Millar, N. and Budgell, B. S. (2019). The passive voice and comprehensibility of biomedical texts: An experimental study with 2 cohorts of chiropractic students. *Journal of Chiropractic Education*, 33(1):16–20.

Mitchell, S. (2007). Medical writing: A prescription for clarity.

Mobiny, A., Yuan, P., Moulik, S. K., Garg, N., Wu, C. C., and Van Nguyen, H. (2021). Dropconnect is effective in modeling uncertainty of bayesian deep networks. *Scientific reports*, 11(1):5458.

Moharasan, G. and Ho, T.-B. (2019). Extraction of temporal information from clinical narratives. *Journal of Healthcare Informatics Research*, 3(2):220–244.

Moiseev, I., Balabaeva, K., and Kovalchuk, S. (2025). Open and extensible benchmark for explainable artificial intelligence methods. *Algorithms*, 18(2):85.

Mondal, I., Hou, Y., and Jochim, C. (2021). End-to-end construction of nlp knowledge graph. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1885–1895.

Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., and Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265.

Morad, G., Helmink, B. A., Sharma, P., and Wargo, J. A. (2021). Hallmarks of response, resistance, and toxicity to immune checkpoint blockade. *Cell*, 184(21):5309–5337.

Moreno-Sánchez, P. (2023). Methods and metrics for evalu-ating explainable artificial in-telligence in healthcare domain. *Bachelor Thesis, Faculty of medicine and health technology, Tampere University.*

Mucsányi, B., Kirchhof, M., and Oh, S. J. (2024). Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks.

Mudelsee, M. (2002). Tauest: a computer program for estimating persistence in unevenly spaced weather/climate time series. *Computers & Geosciences*, 28:69–72.

Murdoch, T. B. and Detsky, A. S. (2013). The inevitable application of big data to health care. *Jama*, 309(13):1351–1352.

Murphy, R. M., Klopotowska, J. E., de Keizer, N. F., Jager, K. J., Leopold, J. H., Dongelmans, D. A., Abu-Hanna, A., and Schut, M. C. (2023). Adverse drug event detection using natural language processing: A scoping review of supervised learning methods. *Plos one*, 18(1):e0279842.

Naeini, M., Cooper, G., and Hauskrecht, M. (2015). Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, volume 2015, pages 2901–2907.

Nakamura, Y., Fujimoto, K., Kluckert, J., Krauthammer, M., Kanzawa, J., Katayama, A., Kikuchi, T., Kurokawa, R., Gonoi, W., Tashiro, Y., Hanaoka, S., Yada, S., Wakamiya, S., and Aramaki, E. (2024). Ntcir-18 radnlp 2024 overview: Dataset and solutions for automated lung cancer staging. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-18. National Institute of Informatics (NII)*.

Nakamura, Y., Hanaoka, S., Yada, S., Wakamiya, S., and Aramaki, E. (2023). Ntcir-17 mednlp-sc radiology report subtask overview: Dataset and solutions for automated lung cancer staging. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-17. National Institute of Informatics (NII)*.

Namata, G., London, B., Getoor, L., Huang, B., and Edu, U. (2012). Query-driven active surveying for collective classification. In *10th international workshop on mining and learning with graphs*, volume 8, page 1.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., Van Keulen, M., and Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42.

Navarro, D. F., Coiera, E., Hambly, T. W., Triplett, Z., Asif, N., Susanto, A., Chowdhury, A., Lorenzo, A. A., Dras, M., and Berkovsky, S. (2025). Expert evaluation of large language models for clinical dialogue summarization. *Scientific Reports*, 15.

Neil, D., Briody, J., Lacoste, A., Sim, A., Creed, P., and Saffari, A. (2018). Interpretable graph convolutional neural networks for inference on noisy knowledge graphs. *arXiv preprint arXiv:1812.00279*.

Neumann, M., King, D., Beltagy, I., and Ammar, W. (2019). ScispaCy: Fast and robust models for biomedical natural language processing. In Demner-Fushman, D., Cohen, K. B., Ananiadou, S., and Tsujii, J., editors, *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Nguyen, D., Rosseel, L., and Grieve, J. (2021). On learning and representing social meaning in nlp: a sociolinguistic perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, pages 603–612.

Nie, Y., Zhou, X., and Bansal, M. (2020a). What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Nie, Y., Zhou, X., and Bansal, M. (2020b). What can we learn from collective human opinions on natural language inference data? In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *EMNLP:2020:main*, pages 9131–9143, Online. acl.

Niu, K., Zhang, K., Peng, X., Pan, Y., and Xiao, N. (2023). Deep multi-modal intermediate fusion of clinical record and time series data in mortality prediction. *Frontiers in Molecular Biosciences*, 10:1136071.

Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., and Tran, D. (2019). Measuring calibration in deep learning. In *CVPR workshops*, volume 2.

nostalgebraist (2020). interpreting GPT: the logit lens.

Obermeyer, Z. and Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine*, 375(13):1216.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

Ogren, P. V., Savova, G., Buntrock, J. D., and Chute, C. G. (2006). Building and evaluating annotated corpora for medical nlp systems. In *AMIA Annual Symposium Proceedings*, volume 2006, page 1050.

Palomaki, J., Rhinehart, O., and Tseng, M. (2018). A case for a range of acceptable annotations. In Aroyo, L., Dumitrache, A., Paritosh, P. K., Quinn, A. J., Welty, C., Checco, A., Demartini, G., Gadiraju, U., and Sarasua, C., editors, *Proceedings of the 1st Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing, and Short Paper Proceedings of the 1st Workshop on Disentangling the Relation Between Crowdsourcing and Bias Management (SAD 2018 and CrowdBias 2018) co-located the 6th AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2018), Zürich, Switzerland, July 5, 2018*, volume 2276 of *CEUR Workshop Proceedings*, pages 19–31. CEUR-WS.org.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318. Pmlr.

Paul, M. J., Sarker, A., Brownstein, J. S., Nikfarjam, A., Scotch, M., Smith, K. L., and Gonzalez, G. (2016). Social media mining for public health monitoring and

surveillance. In *Biocomputing 2016: Proceedings of the Pacific symposium*, pages 468–479. World Scientific.

Paul, S., Mandal, A., Goyal, P., and Ghosh, S. (2023). Pre-trained language models for the legal domain: A case study on indian law. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, ICAIL '23, page 187–196, New York, NY, USA. Association for Computing Machinery.

Pavlick, E. and Kwiatkowski, T. (2019). Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Pedrotti, A., Rambelli, G., Villani, C., and Bolognesi, M. (2025). How humans and LLMs organize conceptual knowledge: Exploring subordinate categories in Italian. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4464–4482, Vienna, Austria. Association for Computational Linguistics.

Peleg, M. (2013). Computer-interpretable clinical guidelines: a methodological review. *Journal of biomedical informatics*, 46(4):744–763.

Peng, Y., Yan, S., and Lu, Z. (2019). Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In Moschitti, A., Pang, B., and Daelemans, W., editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Peters, M. E., Neumann, M., Logan, R., Schwartz, R., Joshi, V., Singh, S., and Smith, N. A. (2019). Knowledge enhanced contextual word representations. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 43–54, Hong Kong, China. Association for Computational Linguistics.

Peterson, J. C., Battleday, R. M., Griffiths, T. L., and Russakovsky, O. (2019). Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Petroni, F., Lewis, P., Piktus, A., Rocktäschel, T., Wu, Y., Miller, A. H., and Riedel, S. (2020). How context affects language models' factual predictions. *arXiv preprint arXiv:2005.04611*.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. (2019). Language models as knowledge bases? In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Pfeiffer, J., Vulić, I., Gurevych, I., and Ruder, S. (2020). Mad-x: An adapter-based framework for multi-task cross-lingual transfer. *arXiv preprint arXiv:2005.00052*.

Pham, T., Tran, T., Phung, D., and Venkatesh, S. (2017). Predicting healthcare trajectories from medical records: A deep learning approach. *Journal of biomedical informatics*, 69:218–229.

Pivovarov, R. and Elhadad, N. (2015). Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, 22(5):938–947.

Piya, F. L. and Beheshti, R. (2025). Contextual: Improving clinical text summarization in llms with context-preserving token filtering and knowledge graphs. *arXiv preprint arXiv:2504.16394*.

Plank, B. (2022). The "problem" of human label variation: On ground truth in data, modeling and evaluation. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *EMNLP:2022:main*, pages 10671–10682, Abu Dhabi, United Arab Emirates. acl.

Plank, B., Hovy, D., and Søgaard, A. (2014). Linguistically debatable or just plain wrong? In Toutanova, K. and Wu, H., editors, *ACL:2014:2*, pages 507–511, Baltimore, Maryland. acl.

Poerner, N., Waltinger, U., and Schütze, H. (2020). Inexpensive domain adaptation of pretrained language models: Case studies on biomedical NER and covid-19 QA. In Cohn, T., He, Y., and Liu, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1482–1490, Online. Association for Computational Linguistics.

Pons, E., Braun, L. M., Hunink, M. M., and Kors, J. A. (2016). Natural language processing in radiology: a systematic review. *Radiology*, 279(2):329–343.

Pradhan, S., Moschitti, A., Xue, N., Ng, H. T., Björkelund, A., Uryupina, O., Zhang, Y., and Zhong, Z. (2013). Towards robust linguistic analysis using OntoNotes. In Hockenmaier, J. and Riedel, S., editors, *Proceedings of the Seventeenth Conference*

*on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

Price, W. N. and Cohen, I. G. (2019). Privacy in the age of medical big data. *Nature medicine*, 25(1):37–43.

Pruthi, G., Liu, F., Kale, S., and Sundararajan, M. (2020). Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33:19920–19930.

Qiu, J., Huang, P., Nakashima, M., Lee, J., Zhu, J., Tang, W., Chen, P., Nguyen, C., Kim, B.-H., Kwon, D., et al. (2023). Multimodal representation learning of cardiovascular magnetic resonance imaging. *arXiv preprint arXiv:2304.07675*.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding with unsupervised learning. *OpenAI Res.*

Raghu, A., Chandak, P., Alam, R., Guttag, J., and Stultz, C. (2022). Contrastive pre-training for multimodal medical time series. In *NeurIPS 2022 Workshop on Learning from Time Series for Health*.

Raj Kanakarajan, K., Kundumani, B., and Sankarasubbu, M. (2021). Bioelectra: pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 143–154.

Rajkomar, A., Dean, J., and Kohane, I. (2018a). Machine learning in medicine. *New England Journal of Medicine*, 378(14):1347–1358.

Rajkomar, A., Dean, J., and Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358.

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., et al. (2018b). Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):1–10.

Rajpurkar, P., Chen, E., Banerjee, O., and Topol, E. J. (2022). Ai in health and medicine. *Nature medicine*, 28(1):31–38.

Rajpurkar, P., Jia, R., and Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. In Gurevych, I. and Miyao, Y., editors, *ACL:2018:2*, pages 784–789, Melbourne, Australia. acl.

Ravuri, S. V., Lenc, K., Willson, M., Kangin, D., Lam, R. R., Mirowski, P. W., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., Prudden, R., Mandhane, A., Clark, A., Brock, A., Simonyan, K., Hadsell, R., Robinson, N. H., Clancy, E., Arribas, A., and Mohamed, S. (2021). Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597:672 – 677.

Rawal, S. (2021). bert-base-uncased-clinical-ner.

Rebholz-Schuhmann, D., Oellrich, A., and Hoehndorf, R. (2012). Text-mining solutions for biomedical research: enabling integrative biology. *Nature Reviews Genetics*, 13(12):829–839.

Rector, A. and Iannone, L. (2012). Lexically suggest, logically define: quality assurance of the use of qualifiers and expected results of post-coordination in snomed ct. *Journal of biomedical informatics*, 45(2):199–209.

Reidsma, D. and Carletta, J. (2008). Squibs: Reliability measurement without limits. *Computational Linguistics*, 34(3):319–326.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *https://arxiv.org/abs/1908.10084*.

Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., Roberts, I., and Setzer, A. (2009). Building a semantically annotated corpus of clinical texts. *Journal of biomedical informatics*, 42(5):950–966.

Roberts, K., Simpson, M., Demner-Fushman, D., Voorhees, E., and Hersh, W. (2016). State-of-the-art in biomedical literature retrieval for clinical cases: a survey of the trec 2014 cds track. *Information Retrieval Journal*, 19(1):113–148.

Rogers, A., Kovaleva, O., and Rumshisky, A. (2021). A primer in bertology: What we know about how bert works. *Transactions of the association for computational linguistics*, 8:842–866.

Romanov, A. and Shivade, C. (2018). Lessons from natural language inference in the clinical domain. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.

Rosenbloom, S. T., Denny, J. C., Xu, H., Lorenzi, N., Stead, W. W., and Johnson, K. B. (2011). Data from clinical notes: a perspective on the tension between structure and flexible documentation. *Journal of the American Medical Informatics Association*, 18(2):181–186.

Ross, A., Marasović, A., and Peters, M. (2021). Explaining NLP models via minimal contrastive editing (MiCE). In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3840–3852, Online. Association for Computational Linguistics.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rumeng, L., Jagannatha Abhyuday, N., and Hong, Y. (2017). A hybrid neural network model for joint prediction of presence and period assertions of medical events in clinical notes. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1149. American Medical Informatics Association.

Sadinle, M., Lei, J., and Wasserman, L. (2019). Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234.

Sahu, G. and Vechtomova, O. (2021). Adaptive fusion techniques for multimodal data. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3156–3166, Online. Association for Computational Linguistics.

Salton, G. (1989). Automatic text processing: The transformation, analysis, and retrieval of. *Reading: Addison-Wesley*, 169.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., and Müller, K.-R. (2016). Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673.

Sánchez, L. G., Zavala, D. E., Farré-Maduell, E., Lima-López, S., Miranda-Escalada, A., and Krallinger, M. (2022). The socialdisner shared task on detection of disease mentions in health-relevant content from social media: methods, evaluation, guidelines and corpora. In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pages 182–189.

Sang, E. F. and De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Sarker, A., Ginn, R., Nikfarjam, A., O'Connor, K., Smith, K., Jayaraman, S., Upadhaya, T., and Gonzalez, G. (2015). Utilizing social media data for pharmacovigilance: a review. *Journal of biomedical informatics*, 54:202–212.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., Tow, J., Rush, A. M., Biderman, S., Webson, A., Ammanamanchi, P. S., Wang, T., Sagot, B., Muennighoff, N., del Moral, A. V., Ruwase, O., Bawden, R., Bekman, S., McMillan-Major, A., Beltagy, I., Nguyen, H., Saulnier, L., Tan, S., Suarez, P. O., Sanh, V., Laurençon, H., Jernite, Y., Launay, J., Mitchell, M., Raffel, C., Gokaslan, A., Simhi, A., Soroa, A., Aji, A. F., Alfassy, A., Rogers, A., Nitzav, A. K., Xu, C., Mou, C., Emezue, C., Klamm, C., Leong, C., van Strien, D., Adelani, D. I., Radev, D., Ponferrada, E. G., Levkovizh, E., Kim, E., Natan, E. B., Toni, F. D., Dupont, G., Kruszewski, G., Pistilli, G., Elsahar, H., Benyamina, H., Tran, H., Yu, I., Abdulmumin, I., Johnson, I., Gonzalez-Dios, I., de la Rosa, J., Chim, J., Dodge, J., Zhu, J., Chang, J., Frohberg, J., Tobing, J., Bhattacharjee, J., Almubarak, K., Chen, K., Lo, K., Werra, L. V., Weber, L., Phan, L., allal, L. B., Tanguy, L., Dey, M., Muñoz, M. R., Masoud, M., Grandury, M., Šaško, M., Huang, M., Coavoux, M., Singh, M., Jiang, M. T.-J., Vu, M. C., Jauhar, M. A., Ghaleb, M., Subramani, N., Kassner, N., Khamis, N., Nguyen, O., Espejel, O., de Gibert, O., Villegas, P., Henderson, P., Colombo, P., Amuok, P., Lhoest, Q., Harliman, R., Bommasani, R., López, R. L., Ribeiro, R.,

Osei, S., Pyysalo, S., Nagel, S., Bose, S., Muhammad, S. H., Sharma, S., Longpre, S., Nikpoor, S., Silberberg, S., Pai, S., Zink, S., Torrent, T. T., Schick, T., Thrush, T., Danchev, V., Nikoulina, V., Laippala, V., Lepercq, V., Prabhu, V., Alyafeai, Z., Talat, Z., Raja, A., Heinzerling, B., Si, C., Taşar, D. E., Salesky, E., Mielke, S. J., Lee, W. Y., Sharma, A., Santilli, A., Chaffin, A., Stiegler, A., Datta, D., Szczechla, E., Chhablani, G., Wang, H., Pandey, H., Strobelt, H., Fries, J. A., Rozen, J., Gao, L., Sutawika, L., Bari, M. S., Al-shaibani, M. S., Manica, M., Nayak, N., Teehan, R., Albanie, S., Shen, S., Ben-David, S., Bach, S. H., Kim, T., Bers, T., Fevry, T., Neeraj, T., Thakker, U., Raunak, V., Tang, X., Yong, Z.-X., Sun, Z., Brody, S., Uri, Y., Tojarieh, H., Roberts, A., Chung, H. W., Tae, J., Phang, J., Press, O., Li, C., Narayanan, D., Bourfoune, H., Casper, J., Rasley, J., Ryabinin, M., Mishra, M., Zhang, M., Shoeybi, M., Peyrounette, M., Patry, N., Tazi, N., Sanseviero, O., von Platen, P., Cornette, P., Lavallée, P. F., Lacroix, R., Rajbhandari, S., Gandhi, S., Smith, S., Requena, S., Patil, S., Dettmers, T., Baruwa, A., Singh, A., Cheveleva, A., Ligozat, A.-L., Subramonian, A., Névéol, A., Lovering, C., Garrette, D., Tunuguntla, D., Reiter, E., Taktasheva, E., Voloshina, E., Bogdanov, E., Winata, G. I., Schoelkopf, H., Kalo, J.-C., Novikova, J., Forde, J. Z., Clive, J., Kasai, J., Kawamura, K., Hazan, L., Carpuat, M., Clinciu, M., Kim, N., Cheng, N., Serikov, O., Antverg, O., van der Wal, O., Zhang, R., Zhang, R., Gehrmann, S., Mirkin, S., Pais, S., Shavrina, T., Scialom, T., Yun, T., Limisiewicz, T., Rieser, V., Protasov, V., Mikhailov, V., Pruksachatkun, Y., Belinkov, Y., Bamberger, Z., Kasner, Z., Rueda, A., Pestana, A., Feizpour, A., Khan, A., Faranak, A., Santos, A., Hevia, A., Unldreaj, A., Aghagol, A., Abdollahi, A., Tammour, A., HajiHosseini, A., Behroozi, B., Ajibade, B., Saxena, B., Ferrandis, C. M., McDuff, D., Contractor, D., Lansky, D., David, D., Kiela, D., Nguyen, D. A., Tan, E., Baylor, E., Ozoani, E., Mirza, F., Ononiwu, F., Rezanejad, H., Jones, H., Bhattacharya, I., Solaiman, I., Sedenko, I., Nejadgholi, I., Passmore, J., Seltzer, J., Sanz, J. B., Dutra, L., Samagaio, M., Elbadri, M., Mieskes, M., Gerchick, M., Akinlolu, M., McKenna, M., Qiu, M., Ghauri, M., Burynok, M., Abrar, N., Rajani, N., Elkott, N., Fahmy, N., Samuel, O., An, R., Kromann, R., Hao, R., Alizadeh, S., Shubber, S., Wang, S., Roy, S., Viguier, S., Le, T., Oyebade, T., Le, T., Yang, Y., Nguyen, Z., Kashyap, A. R., Palasciano, A., Callahan, A., Shukla, A., Miranda-Escalada, A., Singh, A., Beilharz, B., Wang, B., Brito, C., Zhou, C., Jain, C., Xu, C., Fourrier, C., Periñán, D. L., Molano, D., Yu, D., Manjavacas, E., Barth, F., Fuhrimann, F., Altay, G., Bayrak, G., Burns, G., Vrabec, H. U., Bello, I., Dash, I., Kang, J., Giorgi, J., Golde, J., Posada, J. D., Sivaraman, K. R., Bulchandani, L., Liu, L., Shinzato, L., de Bykhovetz, M. H., Takeuchi, M., Pàmies, M., Castillo, M. A., Nezhurina, M., Sänger, M., Samwald, M., Cullan, M., Weinberg, M., Wolf, M. D., Mihaljcic, M., Liu, M., Freidank, M., Kang, M., Seelam, N., Dahlberg, N., Broad, N. M., Muellner, N., Fung, P., Haller, P., Chandrasekhar, R., Eisenberg, R., Martin, R., Canalli, R., Su, R., Su, R., Cahyawijaya, S., Garda, S., Deshmukh, S. S., Mishra, S., Kiblawi, S., Ott, S., Sang-aroonsiri, S., Kumar, S., Schweter, S., Bharati, S., Laud, T., Gigant, T., Kainuma, T., Kusa, W., Labrak, Y., Bajaj, Y. S., Venkatraman, Y., Xu, Y., Xu, Y., Xu, Y., Tan, Z., Xie, Z., Ye, Z., Bras, M., Belkada, Y., and Wolf, T. (2023). Bloom: A 176b-parameter open-access multilingual language model.

Scherrer, Y. (2020). Tapaco: A corpus of sentential paraphrases for 73 languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association (ELRA).

Schmid, P., Cortés, J., Pusztai, L., McArthur, H. L., Kümmel, S., Bergh, J., Denkert, C., Park, Y.-H., Hui, R., Harbeck, N., Takahashi, M., Foukakis, T., Fasching, P. A., Cardoso, F., Untch, M., Jia, L., Karantza, V., Zhao, J., Aktan, G., Dent, R. A., and O'Shaughnessy, J. (2020). Pembrolizumab for early triple-negative breast cancer. *The New England journal of medicine*, 382 9:810–821.

Schnabel, T., Swaminathan, A., Singh, A., Chandak, N., and Joachims, T. (2016). Recommendations as treatments: Debiasing learning and evaluation. In *international conference on machine learning*, pages 1670–1679. PMLR.

Schopf, T., Braun, D., and Matthes, F. (2022a). Evaluating unsupervised text classification: zero-shot and similarity-based approaches. In *Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval*, pages 6–15.

Schopf, T., Braun, D., and Matthes, F. (2022b). Evaluating unsupervised text classification: Zero-shot and similarity-based approaches. In *Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval*, NLPIR 2022, page 6–15. ACM.

Schopf, T., Braun, D., and Matthes, F. (2023). Evaluating unsupervised text classification: Zero-shot and similarity-based approaches. In *Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval*, NLPIR '22, page 6–15. Association for Computing Machinery.

Schröder, C., Niekler, A., and Potthast, M. (2022). Revisiting uncertainty-based query strategies for active learning with transformers. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *FINDINGS:2022:acl*, pages 2194–2203, Dublin, Ireland. acl.

Sellam, T., Das, D., and Parikh, A. (2020). Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Sendak, M. P., Gao, M., Brajer, N., and Balu, S. (2020a). Machine learning in health care: a critical appraisal of challenges and opportunities. *eGEMs*, 8(1):1–15.

Sendak, M. P., Gao, M., Brajer, N., and Balu, S. (2020b). Real-world integration of a sepsis deep learning technology into routine clinical care: Implementation study. *JMIR Medical Informatics*, 8(7):e15182.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In Erk, K. and Smith, N. A., editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Shah, D. S., Schwartz, H. A., and Hovy, D. (2020a). Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.

Shah, H., Tamuly, K., Raghunathan, A., Jain, P., and Netrapalli, P. (2020b). The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems*, 33:9573–9585.

Shang, J., Ma, T., Xiao, C., and Sun, J. (2019). Pre-training of graph augmented transformers for medication recommendation. *arXiv preprint arXiv:1906.00346*.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Shapley, L. S. (1953). A value for n-person games. In Kuhn, H. W. and Tucker, A. W., editors, *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press, Princeton.

Shickel, B., Tighe, P. J., Bihorac, A., and Rashidi, P. (2017). Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604.

Shivade, C., Raghavan, P., Fosler-Lussier, E., Embi, P. J., Elhadad, N., Johnson, S. B., and Lai, A. M. (2014). A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association*, 21(2):221–230.

Shoemake, K. (1985). Animating rotation with quaternion curves. In *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, pages 245–254.

Shukla, S. N. and Marlin, B. (2018). Interpolation-prediction networks for irregularly sampled time series. In *International Conference on Learning Representations*.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. (2023a). Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., Hou, L., Clark, K., Pfohl, S. R., Cole-Lewis, H., et al. (2025). Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S. R., Cole-Lewis, H. J., Neal, D., Schaekermann, M., Wang, A., Amin, M., Lachgar, S., Mansfield, P. A., Prakash, S., Green, B., Dominowska, E., y Arcas, B. A., Tomašev, N., Liu, Y., Wong, R. C., Semturs, C., Mahdavi, S. S., Barral, J. K., Webster, D. R., Corrado, G. S., Matias, Y., Azizi, S., Karthikesalingam, A., and Natarajan, V. (2023b). Towards expert-level medical question answering with large language models. *ArXiv*, abs/2305.09617.

Sinha, A., Ollinger, S., and Constant, M. (2022). Word sense disambiguation of French lexicographical examples using lexical networks. In Ustalov, D., Gao, Y., Panchenko, A., Valentino, M., Thayaparan, M., Nguyen, T. H., Penn, G., Ramesh, A., and Jana, A., editors, *Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing*, pages 70–76, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Soenksen, L. R., Ma, Y., Zeng, C., Boussioux, L., Villalobos Carballo, K., Na, L., Wiberg, H. M., Li, M. L., Fuentes, I., and Bertsimas, D. (2022). Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ digital medicine*, 5(1):149.

Sohn, S., Murphy, S. P., Masanz, J. J., Kocher, J.-P. A., and Savova, G. K. (2010). Classification of medication status change in clinical narratives. In *AMIA Annual Symposium Proceedings*, volume 2010, page 762. American Medical Informatics Association.

Sousa, R. T., Silva, S., Paulheim, H., and Pesquita, C. (2023). Biomedical knowledge graph embeddings with negative statements. In *International Semantic Web Conference*, pages 428–446. Springer.

Spasić, I., Livsey, J., Keane, J. A., and Nenadić, G. (2014). Text mining of cancer-related information: review of current status and future directions. *International journal of medical informatics*, 83(9):605–623.

Squires, S., Harkness, E. F., Evans, D. G., and Astley, S. M. (2023). The effect of variable labels on deep learning models trained to predict breast density. *Biomedical Physics & Engineering Express*, 9:035030.

Steinberg, E., Jung, K., Fries, J. A., Corbin, C. K., Pfohl, S. R., and Shah, N. H. (2021). Language models are an effective representation learning technique for electronic health record data. *Journal of biomedical informatics*, 113:103637.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., and Kattan, M. W. (2010). Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*, 21(1):128–138.

Stubbs, A., Kotfila, C., and Uzuner, Ö. (2015). Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task track 1. *Journal of biomedical informatics*, 58:S11–S19.

Styler IV, W. F., Bethard, S., Finan, S., Palmer, M., Pradhan, S., De Groen, P. C., Erickson, B., Miller, T., Lin, C., Savova, G., et al. (2014). Temporal annotation in

the clinical domain. *Transactions of the association for computational linguistics*, 2:143–154.

Su, L., Hu, C., Li, G., and Cao, D. (2020). Msaf: Multimodal split attention fusion. *arXiv preprint arXiv:2012.07175*.

Su, Y., Wang, M., Wang, P., Zheng, C., Liu, Y., and Zeng, X. (2022). Deep learning joint models for extracting entities and relations in biomedical: a survey and comparison. *Briefings in Bioinformatics*, 23(6):bbac342.

Sun, W., Rumshisky, A., and Uzuner, O. (2013a). Annotating temporal information in clinical narratives. *Journal of biomedical informatics*, 46:S5–S12.

Sun, W., Rumshisky, A., and Uzuner, O. (2013b). Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813.

Sundararajan, M. and Najmi, A. (2020). The many shapley values for model explanation. In *International conference on machine learning*, pages 9269–9278. PMLR.

Sutskever, I., Martens, J., Dahl, G., and Hinton, G. (2013). On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. PMLR.

Swayamdipta, S., Schwartz, R., Lourie, N., Wang, Y., Hajishirzi, H., Smith, N. A., and Choi, Y. (2020). Dataset cartography: Mapping and diagnosing datasets with training dynamics. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *EMNLP:2020:main*, pages 9275–9293, Online. acl.

Talmor, A., Herzig, J., Lourie, N., and Berant, J. (2019). CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Tan, Z. and Tian, Y. (2023). Robust explanation for free or at the cost of faithfulness. In *International conference on machine learning*, pages 33534–33562. PMLR.

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., et al. (2024). Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Tedeschi, S., Maiorca, V., Campolungo, N., Cecconi, F., and Navigli, R. (2021). WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Terada, A., Tokunaga, T., and Tanaka, H. (2004). Automatic expansion of abbreviations by using context and character information. *Information processing & management*, 40(1):31–45.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., and Gurevych, I. (2021). Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., and Ting, D. S. W. (2023). Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018). FEVER: a large-scale dataset for fact extraction and VERification. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Todirascu, A., Padó, S., Krisch, J., Kisselew, M., and Heid, U. (2012). French and german corpora for audience-based text type classification. In *LREC*, volume 2012, pages 1591–1597.

Tonekaboni, S., Joshi, S., McCradden, M. D., and Goldenberg, A. (2019). What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*, pages 359–380. PMLR.

Toneva, M., Sordoni, A., Combes, R. T. d., Trischler, A., Bengio, Y., and Gordon, G. J. (2019). An empirical study of example forgetting during deep neural network learning. In *ICLR*.

Touchent, R., Romary, L., and De La Clergerie, E. (2023). CamemBERT-bio : Un modèle de langue français savoureux et meilleur pour la santé. In Servan, C. and Vilnat, A., editors, *18e Conférence en Recherche d'Information et Applications*
*16e Rencontres Jeunes Chercheurs en RI*
*30e Conférence sur le Traitement Automatique des Langues Naturelles*
*25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, pages 323–334, Paris, France. ATALA.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Turc, I., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962v2*.

Tutubalina, E., Miftahutdinov, Z., Nikolenko, S., and Malykh, V. (2018). Medical concept normalization in social media posts with recurrent neural networks. *Journal of biomedical informatics*, 84:93–102.

Uma, A. N., Fornaciari, T., Hovy, D., Paun, S., Plank, B., and Poesio, M. (2021). Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Uzuner, Ö., Goldstein, I., Luo, Y., and Kohane, I. (2008). Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association*, 15(1):14–24.

Uzuner, Ö., Luo, Y., and Szolovits, P. (2007). Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 14(5):550–563.

Uzuner, Ö., Solti, I., and Cadag, E. (2010). Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518.

Uzuner, Ö., South, B. R., Shen, S., and DuVall, S. L. (2011). 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

ValizadehAslani, T., Shi, Y., Ren, P., Wang, J., Zhang, Y., Hu, M., Zhao, L., and Liang, H. (2023). Pharmbert: a domain-specific bert model for drug labels. *Briefings in Bioinformatics*, 24(4):bbad226.

Vasantharajan, C., Tun, K. Z., Thi-Nga, H., Jain, S., Rong, T., and Siong, C. E. (2022). Medbert: A pre-trained language model for biomedical named entity recognition. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1482–1488.

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *NIPS*.

Verma, A., Griffin, A., Dacre, J., and Elder, A. (2016). Exploring cultural and linguistic influences on clinical communication skills: a qualitative study of international medical graduates. *BMC Medical Education*, 16(1):162.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer-Verlag.

Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018a). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Linzen, T., Chrupała, G., and Alishahi, A., editors, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Wang, P., Han, J., Li, C., and Pan, R. (2019). Logic attention based neighborhood aggregation for inductive knowledge graph embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7152–7159.

Wang, X., Gao, T., Zhu, Z., Zhang, Z., Liu, Z., Li, J., and Tang, J. (2021). KEPLER: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

Wang, Y., Steinhubl, S. R., Defilippi, C., Ng, K., Ebadollahi, S., Stewart, W. F., and Byrd, R. J. (2015). Prescription extraction from clinical notes: towards automating emr medication reconciliation. *AMIA Summits on Translational Science Proceedings*, 2015:188.

Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S., et al. (2018b). Clinical information extraction applications: a literature review. *Journal of biomedical informatics*, 77:34–49.

Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., Gallagher, A., Biswas, R., Ladhak, F., Aarsen, T., et al. (2024). Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

Wei, Y., Peng, J., He, T., Xu, C., Zhang, J., Pan, S., and Chen, S. (2023). Compatible transformer for irregularly sampled multivariate time series. In *2023 IEEE International Conference on Data Mining (ICDM)*, pages 1409–1414. IEEE.

Weinshall, D., Cohen, G., and Amir, D. (2018). Curriculum learning by transfer learning: Theory and experiments with deep networks. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5238–5246. PMLR.

Weiskopf, N. G. and Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 20(1):144–151.

Westbury, S. K., Turro, E., Greene, D., Lentaigne, C., Kelly, A. M., Bariana, T. K., Simeoni, I., Pillois, X., Attwood, A., Austin, S., et al. (2015). Human phenotype ontology annotation and cluster analysis to unravel genetic defects in 707 cases with unexplained bleeding and platelet disorders. *Genome medicine*, 7(1):36.

Wicentowski, R. and Sydes, M. R. (2008). Using implicit information to identify smoking status in smoke-blind medical discharge summaries. *Journal of the American Medical Informatics Association*, 15(1):29–31.

Wiese, G., Weissenborn, D., and Neves, M. (2017). Neural domain adaptation for biomedical question answering. *arXiv preprint arXiv:1706.03610*.

Williams, A., Nangia, N., and Bowman, S. (2018a). A broad-coverage challenge corpus for sentence understanding through inference. In Walker, M., Ji, H., and Stent, A., editors, *NAACL:2018:1*, pages 1112–1122, New Orleans, Louisiana. acl.

Williams, A., Nangia, N., and Bowman, S. R. (2018b). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL-HLT*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In Liu, Q. and Schlangen, D., editors, *EMNLP:2020:demos*, pages 38–45, Online. acl.

Workum, J. D., Volkers, B. W. S., van de Sande, D., Arora, S., Goeijenbier, M., Gommers, D. A., and van Genderen, M. E. (2025). Comparative evaluation and performance of large language models on expert level critical care questions: a benchmark study. *Critical Care*, 29.

Wu, L., Chen, Y., Shen, K., Guo, X., Gao, H., Li, S., Pei, J., Long, B., et al. (2023). Graph neural networks for natural language processing: A survey. *Foundations and Trends® in Machine Learning*, 16(2):119–328.

Wu, T., Ribeiro, M. T., Heer, J., and Weld, D. (2021a). Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6707–6723, Online. Association for Computational Linguistics.

Wu, Y., Hernández-Lobato, J. M., and Ghahramani, Z. (2013). Dynamic covariance models for multivariate financial time series. In *International Conference on Machine Learning*.

Wu, Y., Ni, J., Cheng, W., Zong, B., Song, D., Chen, Z., Liu, Y., Zhang, X., Chen, H., and Davidson, S. B. (2021b). Dynamic gaussian mixture based deep generative model for robust forecasting on sparse multivariate time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 651–659.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Wu, Y., Xu, J., Jiang, M., Zhang, Y., and Xu, H. (2015). A study of neural word embeddings for named entity recognition in clinical text. In *AMIA annual symposium proceedings*, volume 2015, page 1326.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S. (2019). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32:4–24.

Wu, Z., Pan, S., Long, G., Jiang, J., Chang, X., and Zhang, C. (2020). Connecting the dots: Multivariate time series forecasting with graph neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 753–763.

Xiao, J., Basso, L., Nejdl, W., Ganguly, N., and Sikdar, S. (2024). Ivp-vae: Modeling ehr time series with initial value problem solvers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16023–16031.

Xiao, Y., Liang, P. P., Bhatt, U., Neiswanger, W., Salakhutdinov, R., and Morency, L.-P. (2022). Uncertainty quantification with pre-trained language models: A large-scale empirical analysis. *arXiv preprint arXiv:2210.04714*.

Xiao, Y. and Wang, W. Y. (2019a). Quantifying uncertainties in natural language processing tasks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7322–7329.

Xiao, Y. and Wang, W. Y. (2019b). Quantifying uncertainties in natural language processing tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7322–7329.

Xu, R., Musen, M. A., and Shah, N. H. (2010). A comprehensive analysis of five million umls metathesaurus terms using eighteen million medline citations. In *AMIA annual symposium proceedings*, volume 2010, page 907.

Xu, Z., Wang, Y., Bai, L., and Cui, L. (2022). Writing style aware document-level event extraction. *arXiv preprint arXiv:2201.03188*.

Yadav, V. and Bethard, S. (2018). A survey on recent advances in named entity recognition from deep learning models. In Bender, E. M., Derczynski, L., and Isabelle, P., editors, *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yan, L. K. Q., Niu, Q., Li, M., Zhang, Y., Yin, C. H., Fei, C., Peng, B., Bi, Z., Feng, P., Chen, K., Wang, T., Wang, Y., Chen, S., Liu, M., and Liu, J. (2024). Large language model benchmarks in medical tasks.

Yang, B. and Wu, L. (2021). How to leverage multimodal ehr data for better medical predictions? *arXiv preprint arXiv:2110.15763*.

Yang, H., Kuang, L., and Xia, F. (2021). Multimodal temporal-clinical note network for mortality prediction. *Journal of Biomedical Semantics*, 12(1):1–14.

Yang, P., Wang, H., Huang, Y., Yang, S., Zhang, Y., Huang, L., Zhang, Y., Wang, G., Yang, S., He, L., et al. (2024). Lmkg: A large-scale and multi-source medical knowledge graph for intelligent medicine applications. *Knowledge-Based Systems*, 284:111323.

Yang, Z., Cohen, W., and Salakhudinov, R. (2016). Revisiting semi-supervised learning with graph embeddings. In *International conference on machine learning*, pages 40–48. PMLR.

Yasunaga, M., Bosselut, A., Ren, H., Zhang, X., Manning, C. D., Liang, P. S., and Leskovec, J. (2022). Deep bidirectional language-knowledge graph pretraining. *Advances in Neural Information Processing Systems*, 35:37309–37323.

Yeganova, L., Kim, W. G., Comeau, D., Wilbur, W. J., and Lu, Z. (2021). Measuring the relative importance of full text sections for information retrieval from scientific literature. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 247–256, Online. Association for Computational Linguistics.

Yim, W.-w., Fu, Y., Ben Abacha, A., Snider, N., Lin, T., and Yetisgen, M. (2023). Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific Data*, 10(1):586.

Yin, Z. and Shen, Y. (2018). On the dimensionality of word embedding. *Advances in neural information processing systems*, 31.

Yu, D. and Vydiswaran, V. V. (2022). An assessment of mentions of adverse drug events on social media with natural language processing: model development and analysis. *JMIR Medical Informatics*, 10(9):e38140.

Yu, L., Hermann, K. M., Blunsom, P., and Pulman, S. (2014). Deep learning for answer sentence selection. *arXiv preprint arXiv:1412.1632*.

Yu, S., Guo, J., Zhang, R., Fan, Y., Wang, Z., and Cheng, X. (2022). A re-balancing strategy for class-imbalanced classification based on instance difficulty. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 70–79.

Yuan, C., Zhao, K., Kuruoglu, E. E., Wang, L., Xu, T., Huang, W., Zhao, D., Cheng, H., and Rong, Y. (2025). A survey of graph transformers: Architectures, theories and applications. *ArXiv*, abs/2502.16533.

Yuksekgonul, M., Zhang, L., Zou, J. Y., and Guestrin, C. (2023). Beyond confidence: Reliable models should also consider atypicality. *Advances in Neural Information Processing Systems*, 36:38420–38453.

Zarin, D. A., Tse, T., Williams, R. J., Califf, R. M., and Ide, N. C. (2011). The clinicaltrials. gov results database—update and key issues. *New England Journal of Medicine*, 364(9):852–860.

Zeng, F. and Gao, W. (2022). Early rumor detection using neural hawkes process with a new benchmark dataset. In *North American Chapter of the Association for Computational Linguistics*.

Zhang, D., Xue, X., Gao, P., Jin, Z., Hu, M., Wu, Y., and Ying, X. (2024a). A survey of datasets in medicine for large language models. *Intelligence & Robotics*, 4(4):457–478.

Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al. (2023a). Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019a). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhang, X., Chowdhury, R. R., Gupta, R. K., and Shang, J. (2024b). Large language models for time series: A survey. *arXiv preprint arXiv:2402.01801*.

Zhang, X., Dou, D., and Wu, J. (2020). Learning conceptual-contextual embeddings for medical text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9579–9586.

Zhang, X., Li, S., Chen, Z., Yan, X., and Petzold, L. R. (2023b). Improving medical predictions by irregular multimodal electronic health records modeling. In *International Conference on Machine Learning*, pages 41300–41313. PMLR.

Zhang, X., Li, S., Chen, Z., Yan, X., and Petzold, L. R. (2023c). Improving medical predictions by irregular multimodal electronic health records modeling. In *International Conference on Machine Learning*, pages 41300–41313. PMLR.

Zhang, X., Tian, C., Yang, X., Chen, L., Li, Z., and Petzold, L. R. (2023d). Alpacare:instruction-tuned large language models for medical application.

Zhang, X., Zeman, M., Tsiligkaridis, T., and Zitnik, M. (2021). Graph-guided network for irregularly sampled multivariate time series. In *International Conference on Learning Representations*.

Zhang, Y., Baldridge, J., and He, L. (2019b). Paws: Paraphrase adversaries from word scrambling. *arXiv preprint arXiv:1904.01130*.

Zhang, Y., Chen, Q., Yang, Z., Lin, H., and Lu, Z. (2019c). Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific data*, 6(1):52.

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., et al. (2023e). Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Zhang, Y., Rahman, M. M., Braylan, A., Dang, B., Chang, H.-L., Kim, H., McNamara, Q., Angert, A., Banner, E., Khetan, V., et al. (2016). Neural information retrieval: A literature review. *arXiv preprint arXiv:1611.06792*.

Zhang, Y. and Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*.

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., and Liu, Q. (2019d). ERNIE: Enhanced language representation with informative entities. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Zhang, Z., Strubell, E., and Hovy, E. (2022). A survey of active learning for natural language processing. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *EMNLP:2022:main*, pages 6166–6190, Abu Dhabi, United Arab Emirates. acl.

Zhou, X., Liu, H., Pourpanah, F., Zeng, T., and Wang, X. (2022). A survey on epistemic (model) uncertainty in supervised learning: Recent advances and applications. *Neurocomputing*, 489:449–465.

Zhou, Y., Yan, Y., Han, R., Caufield, J. H., Chang, K.-W., Sun, Y., Ping, P., and Wang, W. (2021). Clinical temporal relation extraction with probabilistic soft logic regularization and global inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14647–14655.

Zhu, Y., Li, H., Liao, Y., Wang, B., Guan, Z., Liu, H., and Cai, D. (2017). What to do next: Modeling user behaviors by time-lstm. In *IJCAI*, volume 17, pages 3602–3608.

# Appendix A

# Annotation Guidelines



This appendix outlines the annotation guidelines used for dataset labeling to ensure consistency and reproducibility. Section A.1 describes the specific protocol followed for annotating the SST2 dataset, including labeling criteria and examples where applicable.

## A.1 SST2 dataset

We sample 50 instances for which the length of the sentence is greater than 5. Then, for each instance, we select two SHAP interpretations, one from the model having highest $\mathbb{S}_{interpretation}$ and one from the model having the lowest $\mathbb{S}_{interpretation}$. We select three annotators and ask them to answer for each instance the following three questions:

(a) *which of the two explanation is easier?*

(b) *which of the two explanation is appropriate?*

(c) *which of the two explanation is more suitable?*

The detailed description for each question can be seen in Table A.1.

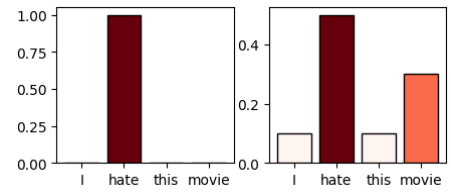
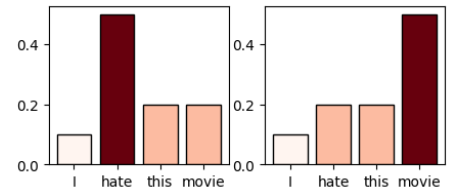
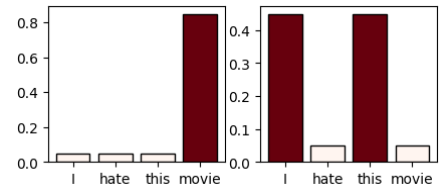
| Criterion | Description | | |
|---|---|---|---|
| Which is simple ? | which of the two explanations is more concise? i.e. Is there an explanation that assigns weights to fewer words? | | |
| Which is appropriate ? | which of the two explanations puts greater emphasis on words you would have paid attention to if you were to classify this sentence? | | |
| Which is easy to make sense of ? | can you reconstruct a reasoning behind weighting these specific words, and if so, which of the explanations corresponds to the most straightforward reasoning process? | | |
| | *simple* | *appropriate* | make sense |
|   | A | A | tie |
|   | tie | A | A |
|   | A | tie | A |

Table A.1: SST-2 Annotation Guidelines

# Appendix B

# Reproducibility



This appendix provides supplementary information to support the main chapters. Appendix B.1 outlines detailed dataset descriptions used across various experiments. Appendix B.2 presents extended details for selected case studies and analyses that were referenced in the main text but abbreviated for brevity. Finally, Appendix B.3 lists relevant code repositories to facilitate reproducibility and further exploration.

## B.1 Datasets Details

This section describes the various dataset and their associated language understanding tasks that we use in our study.

**SocialDisNER** (Sánchez et al., 2022) is a corpus of Spanish Twitter posts annotated for disease mentions. Posts in this dataset originate from a diverse range of users: (i) patients reporting firsthand health experiences, (ii) friends, relatives, and members of support networks sharing the challenges faced by patients, and (iii) medical professionals disseminating authoritative information about diseases.

**CMED** (Mahajan et al., 2020) is a dataset derived from the 2014 i2b2/UTHealth Natural Language Processing shared task corpus (Kumar et al., 2015; Stubbs et al., 2015). CMED comprises a total of 500 clinical notes, partitioned into 350 for training, 50 for development, and 100 for testing.

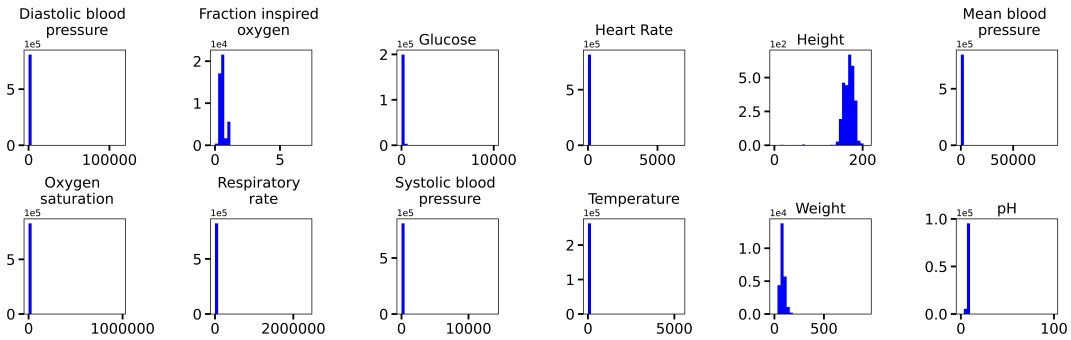
**RefoMED** (Buhnila and Todirascu, 2023) is a collection of French medical terms and corresponding sub-sentential paraphrases. The dataset was built by automatically extracting sentences from the following source corpora, namely ClassYN (Todirascu et al., 2012) and CLEAR Cochrane (Grabar and Cardon, 2018). The paraphrases were identified with the help of linguistic paraphrase markers such as *c'est-à-dire* ("so called"), *également appelé* ("also called"), *est une maladie* ("is a disease"), and punctuation signs, such as colons and brackets. The dataset is made of 6297 pairs of unique medical terms and their corresponding sub-sentential paraphrases. The shortest paraphrase is of 1 word length whereas the longest is 83 words length. The mean and standard deviation are 10.34 and 8.15 respectively.

**PUBMED** (Namata et al., 2012) dataset contains 19717 articles belonging to 3 classes of diabetes-mellitus, `Experimental`, `Type-1`, and `Type-2`.

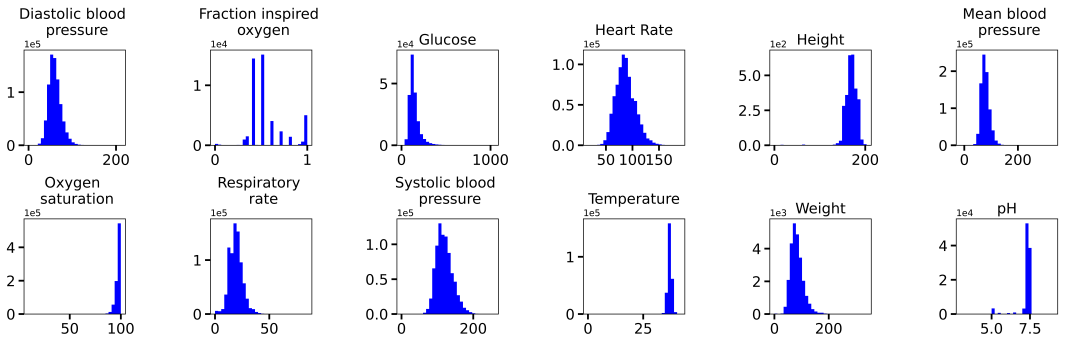
**RADNLP** (Nakamura et al., 2024). It contains two languages : en and jp, where the en version of the dataset in a machine translated version of jp dataset. It is a collection of 378 radiology reports generated from the interpretations of 42 lung cancer cases by nine different board-certified radiologists annotated for TNM Clinical Staging. Of these, 27 cases (243 reports) were created and utilized in NTCIR-17, while 15 cases (135 reports) were originally developed for NTCIR-16.

**INSOMNIA** dataset is a subset csollection of electronic health records (EHRs) from MIMIC-III Clinical Database (v1.4) Johnson et al. (2016) that was annotated by experts of symptoms of Insomnia.

**MIMIC III** It is a dataset of stays of patients in the critical care unit at a large tertiary care hospital. It has 21142 stays of unique patients (instances) with a median length of stay of 2.1 days. A total of 17 physiological measurements, like vital signs, medications, etc., are recorded for each patient. Following SeFT Horn et al. (2020), we remove 32 instances. The discarded instances contained dramatically different recording frequencies compared to the rest of the dataset. Thus, the total number of instances is 21110. We train our model for the in-hospital mortality prediction tasks. Some of the features with numerical data type have extreme outlier values, like oxygen saturation, which should have values in the range of 0-100, but some values are in the range of $10^5$ (see Figure B.1a), possibly due to input/formatting error. Therefore, we remove these outliers. From the training data, 0.008% extreme values are removed in each numerical feature. 0.008% is selected based on the histogram chart of each feature in the training data, as it does not cause too much loss of information and forms a well-distributed histogram, as shown in Figure B.1b. Based on the lower and upper bound values with respect to 0.008% extreme values, the outliers from the test and validation data are also removed.




(a) Histogram of the numerical features of the MIMIC-III dataset before outlier removal.




(b) Histogram of the numerical features of the MIMIC-III dataset after outlier removal.

Figure B.1: Change in the distribution of numerical features in MIMIC-III dataset after removing 0.0008% extreme outlier values.

**Physionet 2012** is a dataset of 12000 patient records (instances) containing measurements taken during the first 48 hours of the ICU stays. Each instance is associated with 37 time series variables (sensors) like blood pressure, lactate, respiration rate, etc., and 6 static descriptor features (i.e., RecordID, Age, Gender, height, ICUType, and

Weight). We follow the SeFT Horn et al. (2020) paper and remove 12 instances that do not contain any time series information. The weight feature is considered a time series since it is measured multiple times in the observation period. The final dataset has 11988 instances with 37 features. We train our model on the in-hospital mortality task, which is a binary classification task to predict if the patient dies before being discharged by using the data of the first 48 hours of the ICU admission.

**SST-2** This is a sentence-level dataset where each sentence is associated with one of the two labels: `positive` or `negative`.

**ChaosNLI** This dataset is collection of subset of three natural language inference datasets (SNLI, MNLI, ANLI). In general, a sample contains a premise and hypothesis and there can be 3 labels: `neutral`, `contradiction` and `entailment`. Exception to the aforementioned description is ANLI, where a sample contains two observation and two hypothesis. In this study, we drop ANLI to keep our experiment setup simple.

**HateXplain** This dataset is also sentence-level. The class labels can be `toxic`, `offensive` and `normal`. The dataset is annotated with token-level input attribution.

**MedABS** The dataset is paragraph level classification. It is collection of medical abstracts describing 5 different classes of patient condition as follows : `Neoplasms`, `Digestive system`, `Nervous system`, `Cardiovascular`, and `General pathological condition`.

**MedNLI** The dataset in medical language inference dataset. It contains a premise and hypothesis pair, which are associated with one of the three possible labels : `neutral`, `contradiction` and `entailment`.

## B.2 Supplementary Details

### Case Study I: Disease identification in tweets

| Model | Identifier |
|---|---|
| ES | flair-nlp/classic-word-embeddings |
| ES+CLIN | flair-nlp/flair-embeddings |
| ES+EN+CLIN | flair-nlp/flair-embeddings |
| BBUCN | bert-base-uncased-clinical-ner |
| BSCFN | BETO |
| XRL | XLM-R |
| BBMCN | bert-base-multilingual-cased-nerhrl |
| WMN | wikineural-multilingual-ner |
| XLRSC | (CLIN-$X_{ES}$) |
| RBBCE | bio-cli-52k |
| SDF | Spanish-disease-finder |

Table B.1: Model Ids for different embedding models.

# Case Study V: Impact of writing style variation

| | Key | Feature-name |
|---|---|---|
| 0 | t_word | total_number_of_words |
| 1 | t_stopword | total_number_of_stop_words |
| 2 | t_punct | total_number_of_puntuations |
| 3 | t_syll | total_number_of_syllables |
| 4 | t_syll2 | total_number_of_words_more_than_two_syllables |
| 5 | t_syll3 | total_number_of_words_more_than_three_syllables |
| 6 | t_uword | total_number_of_unique_words |
| 7 | t_sent | total_number_of_sentences |
| 8 | t_char | total_number_of_characters |
| 9 | a_word_ps | average_number_of_words_per_sentence |
| 10 | a_char_ps | average_number_of_characters_per_sentence |
| 11 | a_char_pw | average_number_of_characters_per_word |
| 12 | a_syll_ps | average_number_of_syllables_per_sentence |
| 13 | a_syll_pw | average_number_of_syllables_per_word |
| 14 | a_stopword_ps | average_number_of_stop_words_per_sentence |
| 15 | a_stopword_pw | average_number_of_stop_words_per_word |
| 16 | t_kup | total_kuperman_age_of_acquistion_of_words |
| 17 | t_bry | total_brysbaert_age_of_acquistion_of_words |
| 18 | t_subtlex_us_zipf | total_subtlex_us_zipf_of_words |
| 19 | a_kup_pw | average_kuperman_age_of_acquistion_of_words_pe... |
| 20 | a_bry_pw | average_brysbaert_age_of_acquistion_of_words_p... |
| 21 | a_kup_ps | average_kuperman_age_of_acquistion_of_words_pe... |
| 22 | a_bry_ps | average_brysbaert_age_of_acquistion_of_words_p... |
| 23 | a_subtlex_us_zipf_pw | average_subtlex_us_zipf_of_words_per_word |
| 24 | a_subtlex_us_zipf_ps | average_subtlex_us_zipf_of_words_per_sentence |
| 25 | t_n_ent | total_number_of_named_entities |
| 26 | t_n_ent_person | total_number_of_named_entities_person |
| 27 | t_n_ent_norp | total_number_of_named_entities_norp |
| 28 | t_n_ent_fac | total_number_of_named_entities_fac |
| 29 | t_n_ent_org | total_number_of_named_entities_org |
| 30 | t_n_ent_gpe | total_number_of_named_entities_gpe |
| 31 | t_n_ent_loc | total_number_of_named_entities_loc |
| 32 | t_n_ent_product | total_number_of_named_entities_product |
| 33 | t_n_ent_event | total_number_of_named_entities_event |
| 34 | t_n_ent_art | total_number_of_named_entities_art |
| 35 | t_n_ent_law | total_number_of_named_entities_law |
| 36 | t_n_ent_language | total_number_of_named_entities_language |
| 37 | t_n_ent_date | total_number_of_named_entities_date |
| 38 | t_n_ent_time | total_number_of_named_entities_time |
| 39 | t_n_ent_percent | total_number_of_named_entities_percent |
| 40 | t_n_ent_money | total_number_of_named_entities_money |
| 41 | t_n_ent_quantity | total_number_of_named_entities_quantity |
| 42 | t_n_ent_ordinal | total_number_of_named_entities_ordinal |
| 43 | t_n_ent_cardinal | total_number_of_named_entities_cardinal |
| 44 | a_n_ent_pw | average_number_of_named_entities_per_word |
| 45 | a_n_ent_person_pw | average_number_of_named_entities_person_per_word |
| 46 | a_n_ent_norp_pw | average_number_of_named_entities_norp_per_word |
| 47 | a_n_ent_fac_pw | average_number_of_named_entities_fac_per_word |
| 48 | a_n_ent_org_pw | average_number_of_named_entities_org_per_word |
| 49 | a_n_ent_gpe_pw | average_number_of_named_entities_gpe_per_word |
| 50 | a_n_ent_loc_pw | average_number_of_named_entities_loc_per_word |
| 51 | a_n_ent_product_pw | average_number_of_named_entities_product_per_word |
| 52 | a_n_ent_event_pw | average_number_of_named_entities_event_per_word |
| 53 | a_n_ent_art_pw | average_number_of_named_entities_art_per_word |
| 54 | a_n_ent_law_pw | average_number_of_named_entities_law_per_word |

| | Key | Feature-name |
|---|---|---|
| 55 | a_n_ent_language_pw | average_number_of_named_entities_language_per_... |
| 56 | a_n_ent_date_pw | average_number_of_named_entities_date_per_word |
| 57 | a_n_ent_time_pw | average_number_of_named_entities_time_per_word |
| 58 | a_n_ent_percent_pw | average_number_of_named_entities_percent_per_word |
| 59 | a_n_ent_money_pw | average_number_of_named_entities_money_per_word |
| 60 | a_n_ent_quantity_pw | average_number_of_named_entities_quantity_per_... |
| 61 | a_n_ent_ordinal_pw | average_number_of_named_entities_ordinal_per_word |
| 62 | a_n_ent_cardinal_pw | average_number_of_named_entities_cardinal_per_... |
| 63 | a_n_ent_ps | average_number_of_named_entities_per_sentence |
| 64 | a_n_ent_person_ps | average_number_of_named_entities_person_per_se... |
| 65 | a_n_ent_norp_ps | average_number_of_named_entities_norp_per_sent... |
| 66 | a_n_ent_fac_ps | average_number_of_named_entities_fac_per_sentence |
| 67 | a_n_ent_org_ps | average_number_of_named_entities_org_per_sentence |
| 68 | a_n_ent_gpe_ps | average_number_of_named_entities_gpe_per_sentence |
| 69 | a_n_ent_loc_ps | average_number_of_named_entities_loc_per_sentence |
| 70 | a_n_ent_product_ps | average_number_of_named_entities_product_per_s... |
| 71 | a_n_ent_event_ps | average_number_of_named_entities_event_per_sen... |
| 72 | a_n_ent_art_ps | average_number_of_named_entities_art_per_sentence |
| 73 | a_n_ent_law_ps | average_number_of_named_entities_law_per_sentence |
| 74 | a_n_ent_language_ps | average_number_of_named_entities_language_per_... |
| 75 | a_n_ent_date_ps | average_number_of_named_entities_date_per_sent... |
| 76 | a_n_ent_time_ps | average_number_of_named_entities_time_per_sent... |
| 77 | a_n_ent_percent_ps | average_number_of_named_entities_percent_per_s... |
| 78 | a_n_ent_money_ps | average_number_of_named_entities_money_per_sen... |
| 79 | a_n_ent_quantity_ps | average_number_of_named_entities_quantity_per_... |
| 80 | a_n_ent_ordinal_ps | average_number_of_named_entities_ordinal_per_s... |
| 81 | a_n_ent_cardinal_ps | average_number_of_named_entities_cardinal_per_... |
| 82 | simp_adj_var | simple_adjectives_variation |
| 83 | simp_adp_var | simple_adpositions_variation |
| 84 | simp_adv_var | simple_adverbs_variation |
| 85 | simp_aux_var | simple_auxiliaries_variation |
| 86 | simp_cconj_var | simple_coordinating_conjunctions_variation |
| 87 | simp_det_var | simple_determiners_variation |
| 88 | simp_intj_var | simple_interjections_variation |
| 89 | simp_noun_var | simple_nouns_variation |
| 90 | simp_num_var | simple_numerals_variation |
| 91 | simp_part_var | simple_particles_variation |
| 92 | simp_pron_var | simple_pronouns_variation |
| 93 | simp_propn_var | simple_proper_nouns_variation |
| 94 | simp_punct_var | simple_punctuations_variation |
| 95 | simp_sconj_var | simple_subordinating_conjunctions_variation |
| 96 | simp_sym_var | simple_symbols_variation |
| 97 | simp_verb_var | simple_verbs_variation |
| 98 | simp_space_var | simple_spaces_variation |
| 99 | root_adj_var | root_adjectives_variation |
| 100 | root_adp_var | root_adpositions_variation |
| 101 | root_adv_var | root_adverbs_variation |
| 102 | root_aux_var | root_auxiliaries_variation |
| 103 | root_cconj_var | root_coordinating_conjunctions_variation |
| 104 | root_det_var | root_determiners_variation |
| 105 | root_intj_var | root_interjections_variation |
| 106 | root_noun_var | root_nouns_variation |
| 107 | root_num_var | root_numerals_variation |
| 108 | root_part_var | root_particles_variation |
| 109 | root_pron_var | root_pronouns_variation |

Table B.2: Linguistic Features (0-109)

| | Key | Feature-name |
|---|---|---|
| 110 | root_propn_var | root_proper_nouns_variation |
| 111 | root_punct_var | root_punctuations_variation |
| 112 | root_sconj_var | root_subordinating_conjunctions_variation |
| 113 | root_sym_var | root_symbols_variation |
| 114 | root_verb_var | root_verbs_variation |
| 115 | root_space_var | root_spaces_variation |
| 116 | corr_adj_var | corrected_adjectives_variation |
| 117 | corr_adp_var | corrected_adpositions_variation |
| 118 | corr_adv_var | corrected_adverbs_variation |
| 119 | corr_aux_var | corrected_auxiliaries_variation |
| 120 | corr_cconj_var | corrected_coordinating_conjunctions_variation |
| 121 | corr_det_var | corrected_determiners_variation |
| 122 | corr_intj_var | corrected_interjections_variation |
| 123 | corr_noun_var | corrected_nouns_variation |
| 124 | corr_num_var | corrected_numerals_variation |
| 125 | corr_part_var | corrected_particles_variation |
| 126 | corr_pron_var | corrected_pronouns_variation |
| 127 | corr_propn_var | corrected_proper_nouns_variation |
| 128 | corr_punct_var | corrected_punctuations_variation |
| 129 | corr_sconj_var | corrected_subordinating_conjunctions_variation |
| 130 | corr_sym_var | corrected_symbols_variation |
| 131 | corr_verb_var | corrected_verbs_variation |
| 132 | corr_space_var | corrected_spaces_variation |
| 133 | simp_ttr | simple_type_token_ratio |
| 134 | root_ttr | root_type_token_ratio |
| 135 | corr_ttr | corrected_type_token_ratio |
| 136 | bilog_ttr | bilogarithmic_type_token_ratio |
| 137 | uber_ttr | uber_type_token_ratio |
| 138 | simp_ttr_no_lem | simple_type_token_ratio_no_lemma |
| 139 | root_ttr_no_lem | root_type_token_ratio_no_lemma |
| 140 | corr_ttr_no_lem | corrected_type_token_ratio_no_lemma |
| 141 | bilog_ttr_no_lem | bilogarithmic_type_token_ratio_no_lemma |
| 142 | uber_ttr_no_lem | uber_type_token_ratio_no_lemma |
| 143 | n_adj | total_number_of_adjectives |
| 144 | n_adp | total_number_of_adpositions |
| 145 | n_adv | total_number_of_adverbs |
| 146 | n_aux | total_number_of_auxiliaries |
| 147 | n_cconj | total_number_of_coordinating_conjunctions |
| 148 | n_det | total_number_of_determiners |
| 149 | n_intj | total_number_of_interjections |
| 150 | n_noun | total_number_of_nouns |
| 151 | n_num | total_number_of_numerals |
| 152 | n_part | total_number_of_particles |
| 153 | n_pron | total_number_of_pronouns |
| 154 | n_propn | total_number_of_proper_nouns |
| 155 | n_punct | total_number_of_punctuations |
| 156 | n_sconj | total_number_of_subordinating_conjunctions |
| 157 | n_sym | total_number_of_symbols |
| 158 | n_verb | total_number_of_verbs |
| 159 | n_space | total_number_of_spaces |
| 160 | n_uadj | total_number_of_unique_adjectives |
| 161 | n_uadp | total_number_of_unique_adpositions |
| 162 | n_uadv | total_number_of_unique_adverbs |
| 163 | n_uaux | total_number_of_unique_auxiliaries |
| 164 | n_ucconj | total_number_of_unique_coordinating_conjunctions |
| 165 | n_udet | total_number_of_unique_determiners |

| | Key | Feature-name |
|---|---|---|
| 166 | n_uintj | total_number_of_unique_interjections |
| 167 | n_unoun | total_number_of_unique_nouns |
| 168 | n_unum | total_number_of_unique_numerals |
| 169 | n_upart | total_number_of_unique_particles |
| 170 | n_upron | total_number_of_unique_pronouns |
| 171 | n_upropn | total_number_of_unique_proper_nouns |
| 172 | n_upunct | total_number_of_unique_punctuations |
| 173 | n_usconj | total_number_of_unique_subordinating_conjunctions |
| 174 | n_usym | total_number_of_unique_symbols |
| 175 | n_uverb | total_number_of_unique_verbs |
| 176 | n_uspace | total_number_of_unique_spaces |
| 177 | a_adj_pw | average_number_of_adjectives_per_word |
| 178 | a_adp_pw | average_number_of_adpositions_per_word |
| 179 | a_adv_pw | average_number_of_adverbs_per_word |
| 180 | a_aux_pw | average_number_of_auxiliaries_per_word |
| 181 | a_cconj_pw | average_number_of_coordinating_conjunctions_pe... |
| 182 | a_det_pw | average_number_of_determiners_per_word |
| 183 | a_intj_pw | average_number_of_interjections_per_word |
| 184 | a_noun_pw | average_number_of_nouns_per_word |
| 185 | a_num_pw | average_number_of_numerals_per_word |
| 186 | a_part_pw | average_number_of_particles_per_word |
| 187 | a_pron_pw | average_number_of_pronouns_per_word |
| 188 | a_propn_pw | average_number_of_proper_nouns_per_word |
| 189 | a_punct_pw | average_number_of_punctuations_per_word |
| 190 | a_sconj_pw | average_number_of_subordinating_conjunctions_p... |
| 191 | a_sym_pw | average_number_of_symbols_per_word |
| 192 | a_verb_pw | average_number_of_verbs_per_word |
| 193 | a_space_pw | average_number_of_spaces_per_word |
| 194 | a_adj_ps | average_number_of_adjectives_per_sentence |
| 195 | a_adp_ps | average_number_of_adpositions_per_sentence |
| 196 | a_adv_ps | average_number_of_adverbs_per_sentence |
| 197 | a_aux_ps | average_number_of_auxiliaries_per_sentence |
| 198 | a_cconj_ps | average_number_of_coordinating_conjunctions_pe... |
| 199 | a_det_ps | average_number_of_determiners_per_sentence |
| 200 | a_intj_ps | average_number_of_interjections_per_sentence |
| 201 | a_noun_ps | average_number_of_nouns_per_sentence |
| 202 | a_num_ps | average_number_of_numerals_per_sentence |
| 203 | a_part_ps | average_number_of_particles_per_sentence |
| 204 | a_pron_ps | average_number_of_pronouns_per_sentence |
| 205 | a_propn_ps | average_number_of_proper_nouns_per_sentence |
| 206 | a_punct_ps | average_number_of_punctuations_per_sentence |
| 207 | a_sconj_ps | average_number_of_subordinating_conjunctions_p... |
| 208 | a_sym_ps | average_number_of_symbols_per_sentence |
| 209 | a_verb_ps | average_number_of_verbs_per_sentence |
| 210 | a_space_ps | average_number_of_spaces_per_sentence |
| 211 | fkre | flesch_kincaid_reading_ease |
| 212 | fkgl | flesch_kincaid_grade_level |
| 213 | fogi | gunning_fog_index |
| 214 | smog | smog_index |
| 215 | cole | coleman_liau_index |
| 216 | auto | automated_readability_index |
| 217 | rt_fast | reading_time_for_fast_readers |
| 218 | rt_average | reading_time_for_average_readers |
| 219 | rt_slow | reading_time_for_slow_readers |

Table B.3: Linguistic Features (110-219)

## Case Study VIII: Missingness

**Transformer** (Vaswani et al., 2017) architecture, originally developed for sequence modeling in natural language processing, is based on a self-attention mechanism that enables the model to capture long-range dependencies between elements in a sequence. While the standard Transformer assumes regularly spaced inputs with fixed positional encodings, its attention mechanism makes it naturally extensible to irregular time series when adapted appropriately. For irregular time series, continuous-time embeddings or learned time encodings can replace the original fixed positional encodings, allowing the model to effectively handle non-uniform time intervals. This flexibility, combined with its global receptive field and parallelizability, has made the Transformer a foundational model for various irregular temporal data applications.

**SeFT** or Set Function Transformer (Horn et al., 2020) is designed specifically to process irregularly sampled time series data by viewing each multivariate time series as an unordered set of observed feature-time pairs. Rather than treating the input as a sequence, SeFT applies a permutation-invariant set function using attention to aggregate information across observations. Each input observation is embedded as a tuple consisting of the feature value and its corresponding timestamp, and attention layers are used to compute representations that respect the unordered and irregular nature of the data. This approach allows SeFT to model sparse and missing data patterns natively, making it particularly suited for clinical or sensor-based time series with variable sampling rates and missing modalities.

**Raindrop** (Zhang et al., 2021) introduces a graph-based representation of multivariate irregular time series by modeling each variable-time observation as a node in a spatiotemporal graph. The model captures both temporal dynamics and inter-variable dependencies using attention-based message passing on this graph structure. Temporal edges connect observations of the same variable over time, while spatial edges connect observations across different variables at the same or similar times. This formulation allows Raindrop to operate effectively on irregular and asynchronous measurements without requiring imputation or interpolation. By leveraging the graph attention network (GAT) framework, Raindrop integrates heterogeneous observations in a principled way, offering strong performance on real-world medical datasets with highly sparse and irregular sampling patterns.

**CoFormer** or Compatible Transformer (Wei et al., 2023) addresses the challenge of missing and asynchronous measurements in irregular time series by introducing a compatibility function that modulates attention weights based on feature-specific temporal distances. Unlike standard Transformers, CoFormer explicitly accounts for the pairwise compatibility between observations, capturing the idea that some features influence others differently depending on their temporal proximity. The model decomposes attention into value and compatibility components, where the compatibility term dynamically adjusts the attention scores to reflect both time gaps and feature-level in-

teractions. This makes CoFormer particularly effective for healthcare time series data, where variables are sampled at different frequencies and carry domain-specific temporal dependencies.

**Time Requirement and Scalability to Number of Sensors** The worst-case time complexity of SLAN is given by $O((N/B) * T * K)$, where$N$ is the number of instances, $B$ is the batch size, $T$ is the maximum length of the time series, and $K$ is the time complexity to process a single LSTM. We utilize GPU to run SLAN, and therefore, the time complexity to process a single LSTM is $O(H^2)$, where $H$ is the hidden size of LSTM. For each LSTM, at each timestep, input, and weight tensors are constructed of shape $(F*B, H+1, 1)$ and $(F*B, H, H+1)$, which are then further multiplied using the torch.matmul operation, as weight∗input=output, with output shape of $(F * B, H, 1)$. Here, $F$ is the number of sensors. The first dimension given by $F * B$ is parallelized in GPU, so $F * B$ numbers of matrix multiplication are computed in parallel. The overall time complexity becomes equal to the time complexity to multiply a single matrix since all matrices are computed parallelly. Matrix multiplication of matrices with shape $(H, H + 1)$ with $(H + 1, 1)$ has a time complexity of $O(H^2)$. Since the factor $H^2$ comes from parallelized operation in GPUs, it will, therefore, have a very small constant factor compared to the other part $[(N/B) * T]$, which is processed sequentially. Therefore the worst time complexity of SLAN in GPU is given by $O((N/B) * T * H^2)$. It can be seen that the training time of SLAN is dependent on the #Instances $(N)$ and the #Observations $(T)$. Refer to Table 1 in the Supplementary for a definition of #Instances and #Observations. This is also evident in the training time required for M-3 and P-12. SLAN requires training time of 373.55±4.80 and 183.99±9.45 seconds per epoch (s/ep) for M-3 and P-12, respectively. Therefore, the time required for an instance of M-3 and P-12 is $2.55\times10^{-2}$ and $2.40\times10^{-2}$ s/ep, respectively. M-3 requires slightly more time than P-12 because its #Observations are slightly higher. Note that the time required by SLAN does not depend on the number of sensors, as M-3 has 17 sensors, whereas P-12 has 37. Thus, SLAN is scalable to the number of sensors with regard to time complexity. However, if the number of sensors is very high (say 10k), SLAN might suffer from a storage limit since the model would store parameters corresponding to 10k LSTMs. It should be noted that the datasets used in this paper are standard for ISTS applications and can be considered a good representation of practical applications. Thus, SLAN is sufficiently scalable to handle practical scenarios.

# Case Study X: External KB for Medical Paraphrasing

| Setup | TOKEN | PromptID | Split | -bleu | -bert | -bleurt | -rouge1 | -rouge2 | -rougL | -rougeLsum | -bleu-p1 | -bleu-p2 | -bleu-p3 | -bleu-p4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | no-fine-tuned -BIOM-SLERP | | | | | | | | | | | | | |
| no-fine-tuned -BIOM-SLERP | 25 | p4 | val | 0.0 | 0.6917 | 0.1411 | 0.1854 | 0.0655 | 0.1854 | 0.1587 | 0.0905 | 0.0214 | 0.0068 | 0.0014 |
| no-fine-tuned -BIOM-SLERP | 50 | p4 | val | 0.0 | 0.6806 | 0.1564 | 0.1744 | 0.0558 | 0.1744 | 0.1403 | 0.0677 | 0.0124 | 0.0033 | 0.0007 |
| no-fine-tuned -BIOM-SLERP | 25 | p4 | test | 0.0028 | 0.6936 | 0.1423 | 0.1955 | 0.0702 | 0.1955 | 0.1689 | 0.1127 | 0.0311 | 0.014 | 0.0073 |
| no-fine-tuned -BIOM-SLERP | 50 | p4 | test | 0.003 | 0.6804 | 0.1521 | 0.1793 | 0.0597 | 0.1793 | 0.145 | 0.0803 | 0.0205 | 0.0084 | 0.0042 |
| | no-fine-tuned -BARTHEZ | | | | | | | | | | | | | |
| no-fine-tuned -BARTHEZ | 25 | p4 | val | 0.0 | 0.643 | 0.1213 | 0.084 | 0.0189 | 0.084 | 0.0784 | 0.0361 | 0.0029 | 0.0012 | 0.0002 |
| no-fine-tuned -BARTHEZ | 50 | p4 | val | 0.0 | 0.643 | 0.1213 | 0.084 | 0.0189 | 0.084 | 0.0784 | 0.0361 | 0.0029 | 0.0012 | 0.0002 |
| no-fine-tuned -BARTHEZ | 25 | p4 | test | 0.0 | 0.634 | 0.0996 | 0.0699 | 0.0071 | 0.0699 | 0.0635 | 0.0351 | 0.0018 | 0.0002 | 0.0 |
| no-fine-tuned -BARTHEZ | 50 | p4 | test | 0.0 | 0.634 | 0.0996 | 0.0699 | 0.0071 | 0.0699 | 0.0635 | 0.0351 | 0.0018 | 0.0002 | 0.0 |
| | no-fine-tuned -BIOM-SLERP-pRAGe | | | | | | | | | | | | | |
| E:camem D:biomistral | 25 | p4 | val | 0.0 | 0.6859 | 0.1332 | 0.1418 | 0.047 | 0.1418 | 0.1232 | 0.0737 | 0.0165 | 0.0057 | 0.0003 |
| E:camem D:biomistral | 50 | p4 | val | 0.0003 | 0.6746 | 0.1629 | 0.1357 | 0.0366 | 0.1357 | 0.1096 | 0.0621 | 0.0107 | 0.0032 | 0.0006 |
| E:camem D:biomistral | 25 | p4 | test | 0.0 | 0.6909 | 0.1472 | 0.1741 | 0.0592 | 0.1741 | 0.1494 | 0.0948 | 0.0232 | 0.0094 | 0.0018 |
| E:camem D:biomistral | 50 | p4 | test | 0.0001 | 0.6817 | 0.1587 | 0.1621 | 0.0488 | 0.1621 | 0.1306 | 0.0736 | 0.0156 | 0.0054 | 0.0012 |
| E:drbert D:biomistral | 25 | p4 | val | 0.0 | 0.6738 | 0.138 | 0.0952 | 0.0269 | 0.0952 | 0.083 | 0.0501 | 0.011 | 0.003 | 0.0001 |
| E:drbert D:biomistral | 50 | p4 | val | 0.0 | 0.6616 | 0.1482 | 0.0899 | 0.0211 | 0.0899 | 0.0712 | 0.0415 | 0.0069 | 0.0019 | 0.0003 |
| E:drbert D:biomistral | 25 | p4 | test | 0.0 | 0.6744 | 0.1308 | 0.1086 | 0.0352 | 0.1086 | 0.0913 | 0.0574 | 0.0115 | 0.0037 | 0.0007 |
| E:drbert D:biomistral | 50 | p4 | test | 0.0 | 0.6646 | 0.1363 | 0.1008 | 0.029 | 0.1008 | 0.0796 | 0.0465 | 0.0091 | 0.0033 | 0.0014 |
| | no-fine-tuned orangesum-BARTHEZ-pRAGe | | | | | | | | | | | | | |
| E:camem D:barthez | 25 | p4 | val | 0.0 | 0.6565 | 0.0775 | 0.1385 | 0.0269 | 0.1385 | 0.1106 | 0.0575 | 0.0063 | 0.0022 | 0.0001 |
| E:camem D:barthez | 50 | p4 | val | 0.0 | 0.6593 | 0.1472 | 0.1457 | 0.0391 | 0.1457 | 0.1171 | 0.0568 | 0.007 | 0.0022 | 0.0002 |
| E:camem D:barthez | 25 | p4 | test | 0.0 | 0.6521 | 0.0734 | 0.1187 | 0.0188 | 0.1187 | 0.0987 | 0.0538 | 0.0045 | 0.0011 | 0.0002 |
| E:camem D:barthez | 50 | p4 | test | 0.0 | 0.6525 | 0.1127 | 0.1164 | 0.0179 | 0.1164 | 0.0955 | 0.049 | 0.0038 | 0.0011 | 0.0002 |
| E:drbert D:barthez | 25 | p4 | val | 0.0 | 0.6397 | 0.0258 | 0.1209 | 0.0109 | 0.1209 | 0.0958 | 0.0459 | 0.0033 | 0.0001 | 0.0 |
| E:drbert D:barthez | 50 | p4 | val | 0.0 | 0.6329 | 0.0636 | 0.1124 | 0.0109 | 0.1124 | 0.0889 | 0.0402 | 0.0024 | 0.0002 | 0.0 |
| E:drbert D:barthez | 25 | p4 | test | 0.0 | 0.6414 | 0.0235 | 0.1027 | 0.0047 | 0.1027 | 0.0811 | 0.0446 | 0.0005 | 0.0 | 0.0 |
| E:drbert D:barthez | 50 | p4 | test | 0.0 | 0.6517 | 0.0501 | 0.1096 | 0.0074 | 0.1096 | 0.0864 | 0.048 | 0.0016 | 0.0 | 0.0 |

Table B.4: Evaluation report follows equation 4.1 for each metric (without fine tuning).

| Setup | TOKEN | -bleu | -bert | -bleurt | -rouge1 | -rouge2 | -rougL | -rougeLsum | -bleu-p1 | -bleu-p2 | -bleu-p3 | -bleu-p4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| w/o FINE TUNING | | | | | | | | | | | | |
| **BARTHEZ** | 25 | $0.00_{0.00}$ | $0.63_{0.03}$ | $0.10_{0.10}$ | $0.07_{0.08}$ | $0.01_{0.03}$ | $0.07_{0.08}$ | $0.06_{0.07}$ | $0.04_{0.06}$ | $0.00_{0.01}$ | $0.00_{0.00}$ | $0.00_{0.00}$ |
| | 50 | $0.00_{0.00}$ | $0.63_{0.03}$ | $0.10_{0.10}$ | $0.07_{0.08}$ | $0.01_{0.03}$ | $0.07_{0.08}$ | $0.06_{0.07}$ | $0.04_{0.06}$ | $0.00_{0.01}$ | $0.00_{0.00}$ | $0.00_{0.00}$ |
| **BIOMISTRAL** | 25 | $0.00_{0.02}$ | $0.70_{0.06}$ | $0.15_{0.15}$ | $0.20_{0.16}$ | $0.07_{0.12}$ | $0.20_{0.16}$ | $0.17_{0.14}$ | $0.11_{0.12}$ | $0.03_{0.08}$ | $0.01_{0.06}$ | $0.01_{0.05}$ |
| | 50 | $0.00_{0.03}$ | $0.68_{0.06}$ | $0.16_{0.15}$ | $0.18_{0.13}$ | $0.06_{0.09}$ | $0.18_{0.13}$ | $0.14_{0.11}$ | $0.08_{0.08}$ | $0.02_{0.05}$ | $0.01_{0.03}$ | $0.00_{0.03}$ |
| **CAMEMBERT+BARTHEZ** | 25 | $0.00_{0.00}$ | $0.65_{0.05}$ | $0.07_{0.09}$ | $0.12_{0.10}$ | $0.02_{0.05}$ | $0.12_{0.10}$ | $0.10_{0.08}$ | $0.05_{0.07}$ | $0.00_{0.02}$ | $0.00_{0.01}$ | $0.00_{0.00}$ |
| | 50 | $0.00_{0.00}$ | $0.65_{0.05}$ | $0.11_{0.11}$ | $0.12_{0.10}$ | $0.02_{0.05}$ | $0.12_{0.10}$ | $0.10_{0.08}$ | $0.05_{0.06}$ | $0.00_{0.02}$ | $0.00_{0.01}$ | $0.00_{0.00}$ |
| **DRBERT+BARTHEZ** | 25 | $0.00_{0.00}$ | $0.64_{0.03}$ | $0.02_{0.06}$ | $0.10_{0.09}$ | $0.00_{0.02}$ | $0.10_{0.09}$ | $0.08_{0.06}$ | $0.04_{0.06}$ | $0.00_{0.01}$ | $0.00_{0.00}$ | $0.00_{0.00}$ |
| | 50 | $0.00_{0.00}$ | $0.65_{0.04}$ | $0.05_{0.07}$ | $0.11_{0.09}$ | $0.01_{0.02}$ | $0.11_{0.09}$ | $0.09_{0.07}$ | $0.05_{0.06}$ | $0.00_{0.01}$ | $0.00_{0.00}$ | $0.00_{0.00}$ |
| **CAMEMBERT+BIOMISTRAL** | 25 | $0.01_{0.06}$ | $0.69_{0.06}$ | $0.14_{0.15}$ | $0.19_{0.17}$ | $0.08_{0.14}$ | $0.19_{0.17}$ | $0.17_{0.15}$ | $0.12_{0.14}$ | $0.04_{0.12}$ | $0.02_{0.11}$ | $0.02_{0.10}$ |
| | 50 | $0.00_{0.03}$ | $0.68_{0.06}$ | $0.17_{0.15}$ | $0.18_{0.14}$ | $0.06_{0.10}$ | $0.18_{0.14}$ | $0.15_{0.12}$ | $0.08_{0.09}$ | $0.02_{0.05}$ | $0.01_{0.04}$ | $0.01_{0.04}$ |
| **DRBERT+BIOMISTRAL** | 25 | $0.00_{0.02}$ | $0.69_{0.06}$ | $0.14_{0.15}$ | $0.18_{0.17}$ | $0.07_{0.13}$ | $0.18_{0.17}$ | $0.16_{0.16}$ | $0.11_{0.12}$ | $0.03_{0.08}$ | $0.02_{0.06}$ | $0.01_{0.05}$ |
| | 50 | $0.00_{0.02}$ | $0.68_{0.06}$ | $0.17_{0.15}$ | $0.17_{0.13}$ | $0.05_{0.09}$ | $0.17_{0.13}$ | $0.14_{0.12}$ | $0.08_{0.08}$ | $0.02_{0.05}$ | $0.01_{0.04}$ | $0.00_{0.03}$ |
| w/ FINE TUNING | | | | | | | | | | | | |
| **BARTHEZ★** | 25 | $0.00_{0.00}$ | $0.62_{0.02}$ | $0.05_{0.08}$ | $0.11_{0.08}$ | $0.01_{0.02}$ | $0.11_{0.08}$ | $0.09_{0.06}$ | $0.06_{0.07}$ | $0.00_{0.01}$ | $0.00_{0.00}$ | $0.00_{0.00}$ |
| | 50 | $0.00_{0.01}$ | $0.63_{0.03}$ | $0.09_{0.09}$ | $0.12_{0.08}$ | $0.01_{0.04}$ | $0.12_{0.08}$ | $0.09_{0.06}$ | $0.07_{0.07}$ | $0.01_{0.02}$ | $0.00_{0.02}$ | $0.00_{0.01}$ |
| **BIOMISTRAL★** | 25 | $0.00_{0.00}$ | $0.72_{0.07}$ | $0.15_{0.17}$ | $0.22_{0.17}$ | $0.09_{0.13}$ | $0.22_{0.17}$ | $0.20_{0.16}$ | $0.14_{0.13}$ | $0.04_{0.07}$ | $0.01_{0.04}$ | $0.00_{0.02}$ |
| | 50 | $0.00_{0.00}$ | $0.69_{0.07}$ | $0.16_{0.16}$ | $0.18_{0.13}$ | $0.07_{0.10}$ | $0.18_{0.13}$ | $0.16_{0.12}$ | $0.10_{0.10}$ | $0.02_{0.04}$ | $0.01_{0.02}$ | $0.00_{0.01}$ |
| **CAMEMBERT+BARTHEZ★** | 25 | $0.00_{0.00}$ | $0.65_{0.05}$ | $0.05_{0.09}$ | $0.12_{0.10}$ | $0.02_{0.05}$ | $0.12_{0.10}$ | $0.10_{0.08}$ | $0.06_{0.07}$ | $0.01_{0.02}$ | $0.00_{0.01}$ | $0.00_{0.00}$ |
| | 50 | $0.00_{0.01}$ | $0.64_{0.05}$ | $0.10_{0.10}$ | $0.12_{0.10}$ | $0.02_{0.05}$ | $0.12_{0.10}$ | $0.10_{0.08}$ | $0.06_{0.07}$ | $0.01_{0.02}$ | $0.00_{0.01}$ | $0.00_{0.01}$ |
| **DRBERT+BARTHEZ★** | 25 | $0.00_{0.00}$ | $0.64_{0.03}$ | $0.01_{0.04}$ | $0.13_{0.10}$ | $0.01_{0.03}$ | $0.13_{0.10}$ | $0.10_{0.07}$ | $0.06_{0.07}$ | $0.00_{0.01}$ | $0.00_{0.00}$ | $0.00_{0.00}$ |
| | 50 | $0.00_{0.00}$ | $0.64_{0.04}$ | $0.05_{0.07}$ | $0.12_{0.09}$ | $0.01_{0.03}$ | $0.12_{0.09}$ | $0.09_{0.06}$ | $0.05_{0.06}$ | $0.00_{0.01}$ | $0.00_{0.00}$ | $0.00_{0.00}$ |
| **CAMEMBERT+BIOMISTRAL★** | 25 | $0.00_{0.00}$ | $0.60_{0.04}$ | $0.13_{0.11}$ | $0.09_{0.06}$ | $0.02_{0.03}$ | $0.09_{0.06}$ | $0.07_{0.05}$ | $0.03_{0.03}$ | $0.01_{0.01}$ | $0.00_{0.00}$ | $0.00_{0.00}$ |
| | 50 | $0.00_{0.00}$ | $0.60_{0.05}$ | $0.16_{0.11}$ | $0.09_{0.06}$ | $0.02_{0.03}$ | $0.09_{0.06}$ | $0.07_{0.05}$ | $0.03_{0.03}$ | $0.01_{0.01}$ | $0.00_{0.00}$ | $0.00_{0.00}$ |
| **DRBERT+BIOMISTRAL★** | 25 | $0.00_{0.00}$ | $0.59_{0.04}$ | $0.12_{0.15}$ | $0.08_{0.06}$ | $0.02_{0.02}$ | $0.08_{0.06}$ | $0.07_{0.04}$ | $0.03_{0.03}$ | $0.00_{0.01}$ | $0.00_{0.00}$ | $0.00_{0.00}$ |
| | 50 | $0.00_{0.00}$ | $0.60_{0.04}$ | $0.14_{0.15}$ | $0.08_{0.06}$ | $0.02_{0.02}$ | $0.08_{0.06}$ | $0.07_{0.04}$ | $0.03_{0.02}$ | $0.00_{0.01}$ | $0.00_{0.00}$ | $0.00_{0.00}$ |

Table B.5: Detailed Experiment report on test set following equation 4.1 for each metric. The ★ symbol denotes configurations with finetuned SLM.

| | TOKEN | | Base SLM : BIOMISTRAL NON-FINETUNED | FINETUNED |
|---|---|---|---|---|
| | | | PROMPT + **ostéophyte** | |
| SLM | 25 | fr | Un ostéophyte est une formation osseuse qui pousse à partir d'une articulation ou d | des dépôts de tissu osseux qui se forment sur les bords des os[1 |
| | | en | (An osteophyte is a bony formation that grows from a joint or ) | (deposits of bone tissue that form on the edges of bones[1) |
| | | [TAG] | Cr-R:1; Cr-S:1; Cm-R:0; Cm-S:0; Rd:2. | Cr-R:1; Cr-S:1; Cm-R:1. Cm-S:0; Rd:1. |
| | | | PROMPT + **asthme, mucoviscidose, ventilation mécanique** | |
| | | fr | Asthme: maladie où les airways se ferment et se contractent, faisant du bruit lors | maladies respiratoires chroniques et maladies rares respiratoires (M) |
| | | en | (Asthma: a disease in which the airways close and contract, making noise as) | (chronic respiratory diseases and rare respiratory diseases (M) |
| | | [TAG] | Cr-R:1; Cr-S:0; Cm-R:0; Cm-S:0; Rd:1. | Cr-R:1; Cr-S:1; Cm-R:1; Cm-S:0; Rd:1. |
| | | | PROMPT + **ostéophyte** | |
| | 50 | fr | Un ostéophyte est une formation osseuse qui pousse à partir d'une articulation ou d'une jointure. Il peut être formé en réponse à une blessure ou à une inflammation, mais | des dépôts de tissu osseux qui se forment sur les bords des os [1] [2] [3] [4] [5] [6] [7] |
| | | en | (An osteophyte is a bone formation that grows from a joint or articulation. It may be formed in response to injury or inflammation, but) | (deposits of bone tissue that form on the edges of bones [1] [2] [3] [4] [5] [6] [7]) |
| | | [TAG] | Cr-R:1; Cr-S:1; Cm-R:1; Cm-S:0; Rd:2. | Cr-R:1; Cr-S:1; Cm-R:1; Cm-S:0; Rd:1. |
| | | | PROMPT + **asthme, mucoviscidose, ventilation mécanique** | |
| | | fr | Asthme: maladie où les airways se ferment et se contractent, faisant du bruit lors de l'inspiration et de la respiration. Mucoviscidose: maladie où | maladies respiratoires chroniques et maladies rares respiratoires (MRR) (MRR) (MRR) (MRR) (MRR) (M |
| | | en | (Asthma: a disease in which the airways close and contract, making noise when breathing in and out. Cystic fibrosis: a disease in which) | (chronic respiratory diseases and rare respiratory diseases (MRR) (MRR) (MRR) (MRR) (MRR) (MRR) (M) |
| | | [TAG] | Cr-R:1; Cr-S:0; Cm-R:1; Cm-S:0; Rd:1. | Cr-R:1; Cr-S:1; Cm-R:1; Cm-S:0; Rd:1. |
| | | | pRAGe : BIOMISTRAL | |
| | | | PROMPT + **ostéophyte** | |
| C.BERT | 25 | fr | Un ostéophyte est une formation osseuse qui pousse autour d'une articulation. Il | est une formation osseuse qui se développe sur les os entourant une articulation. |
| | | en | (An osteophyte is a bone formation that grows around a joint. It) | (is a bone formation that develops on the bones surrounding a joint.) |
| | | [TAG] | Cr-R:1; Cr-S:1; Cm-R:1; Cm-S:0; Rd:2. | Cr-R:1; Cr-S:1; Cm-R:1; Cm-S:1; Rd:1. |
| | | | PROMPT + **asthme, mucoviscidose, ventilation mécanique** | |
| | | fr | Asthme: maladie qui fait ressentir des difficultés à respirer, souvent accompagnée de | est une maladie génétique qui affecte les glandes salivaires et les glandes lacry |
| | | en | (Asthma: a disease that makes it difficult to breathe, often accompanied b) | (is a genetic disease that affects the salivary and lacrimal glan) |
| | | [TAG] | Cr-R:1; Cr-S:1; Cm-R:0; Cm-S:0; Rd:1. | Cr-R:1; Cr-S:0; Cm-R:0; Cm-S:0; Rd:1. |
| | | | PROMPT + **ostéophyte** | |
| | 50 | fr | Un ostéophyte est une formation osseuse qui pousse autour d'une articulation. Il peut être causé par une maladie appelée arthrose, qui fait que de la matière osseuse | est une formation osseuse qui se développe sur les os entourant une articulation. Les ostéophytes peuvent être symptomatiques ou asymptomatiques. Lors |
| | | en | (An osteophyte is a bone formation that grows around a joint. It can be caused by a disease called arthritis, which causes bone material to) | (is a bone formation that develops on the bones surrounding a joint. Osteophytes may be symptomatic or asymptomatic. At) |
| | | [TAG] | Cr-R:1; Cr-S:1; Cm-R:1; Cm-S:0; Rd:2. | Cr-R:1; Cr-S:1; Cm-R:1; Cm-S:0; Rd:1. |
| | | | PROMPT + **asthme, mucoviscidose, ventilation mécanique** | |
| | | fr | Asthme: maladie qui fait ressentir des difficultés à respirer, souvent accompagnée de toux et de sifflements. Mucoviscidose: maladie génétique qui affect | est une maladie génétique qui affecte les glandes salivaires et les glandes lacrymales, provoquant une production excessive de mucus. Cette maladie peut également affecter les voies |
| | | en | (Asthma: a disease that makes breathing difficult, often accompanied by coughing and wheezing. Cystic fibrosis: genetic disease that affects) | (is a genetic disorder that affects the salivary and lacrimal glands, causing excessive mucus production. This disease can also affect the [tracks]) |
| | | [TAG] | Cr-R:1; Cr-S:1; Cm-R:1; Cm-S:0; Rd:1. | Cr-R:1; Cr-S:1; Cm-R:1; Cm-S:0; Rd:1. |

Table B.6: Examples of generated answers from the different BioMistral configurations.

| | TOKEN | | pRAGe : BARTHEZ NON-FINETUNED | pRAGe : BARTHEZ FINETUNED |
|---|---|---|---|---|
| | | | PROMPT + **ostéophyte** | |
| C.BERT | 25 | fr | L'arthrose est une affection dégénérative des articulations qui fait que de la matière osseuse est anormalement produite | L'ostéophytose désigne le phénomène d'apparition d'un ostéophyte sur une articulation, lors |
| | | en | (Osteoarthritis is a degenerative joint disorder in which bone material is abnormally produced.) | (Osteophytosis refers to the phenomenon of the appearance of an osteophyte on a joint, when) |
| | | [TAG] | Cr-R:0; Cr-S:0; Cm-R:1; Cm-S:0; Rd:1. | Cr-R:1; Cr-S:1; Cm-R:0; Cm-S:0; Rd:2. |
| | | | PROMPT + **asthme, mucoviscidose, ventilation mécanique** | |
| | | fr | Ces descriptions se focalisent le plus souvent sur les problèmes digestifs, stéatorrhée et troubles pancréatiques, | Notes et références Voir aussi Articles connexes BradypnéeDyspnéePolypnéeT |
| | | en | (These descriptions most often focus on digestive problems, steatorrhea and pancreatic disorders,) | (Notes and references See also Related articles BradypneaDyspneaPolypneaT) |
| | | [TAG] | Cr-R:0; Cr-S:0; Cm-R:0; Cm-S:0; Rd:2. | Cr-R:0; Cr-S:0; Cm-R:0; Cm-S:0; Rd:3. |
| | | | PROMPT + **ostéophyte** | |
| | 50 | fr | L'arthrose est une affection dégénérative des articulations qui fait que de la matière osseuse est anormalement produite à l'extérieur de l'os. | L'ostéophytose est une affection dégénérative des articulations qui fait que de la matière osseuse est anormalement produite à l'extérieur de l'os, au détriment de la membrane blanchâtre fibreuse qui entoure l'os |
| | | en | (Osteoarthritis is a degenerative joint disease in which abnormal bone material is produced outside the bone.) | (Osteophytosis is a degenerative joint disorder in which bone material is abnormally produced on the outside of the bone, to the detriment of the whitish fibrous membrane that surrounds the bone. |
| | | [TAG] | Cr-R:1; Cr-S:0; Cm-R:1; Cm-S:1; Rd:1. | Cr-R:1; Cr-S:1; Cm-R:1; Cm-S:1; Rd:2. |
| | | | PROMPT + **asthme, mucoviscidose, ventilation mécanique** | |
| | | fr | Ces descriptions se focalisent le plus souvent sur les problèmes digestifs, stéatorrhée et troubles pancréatiques. | Notes et références Voir aussi Articles connexes BradypnéeDyspnéePolypnéeTachypnéeSyndrome d'apnées du sommeilApnée (sport) |
| | | en | (These descriptions most often focus on digestive problems, steatorrhea and pancreatic disorders.) | (Notes and references See also Related articles BradypneaDyspneaPolypneaTachypneaSleep apnea syndromeApnea (sport))) |
| | | [TAG] | Cr-R:0; Cr-S:0; Cm-R:1; Cm-S:1; Rd:2. | Cr-R:0; Cr-S:0; Cm-R:0; Cm-S:0; Rd:3. |

Table B.7: Examples of generated answers from BARTHEZ in the CamemBERT pRAGe setup.

## Uncertainty awareness of Medical Models

| Models | | MedABS | | MedNLI | | SMOKING | | PxSLU | | MedMCQA | | MORFITT | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | lr | E | lr | E | lr | E | lr | E | lr | E | lr | E |
| $-\mathcal{D}$ | DNN | 1e-5 | 4 | 5e-6 | 12 | 1e-4 | 15 | 5e-6 | 15 | 5e-6 | 14 | 5e-5 | 15 |
| $-\mathcal{D}$ | DC | 5e-6 | 7 | 1e-5 | 11 | 1e-5 | 15 | 5e-6 | 13 | 5e-6 | 15 | 5e-5 | 13 |
| $-\mathcal{D}$ | MCD | 5e-5 | 5 | 5e-6 | 15 | 5e-5 | 15 | 1e-5 | 14 | 5e-6 | 11 | 5e-5 | 10 |
| $-\mathcal{D}$ | VI | 5e-6 | 7 | 1e-5 | 14 | 5e-6 | 13 | 5e-6 | 14 | 1e-6 | 15 | 5e-5 | 13 |
| $+\mathcal{D}$ | DNN | 1e-5 | 4 | 1e-5 | 14 | 5e-5 | 15 | 1e-5 | 15 | 1e-5 | 10 | 5e-5 | 15 |
| $+\mathcal{D}$ | DC | 5e-5 | 3 | 1e-5 | 13 | 1e-4 | 13 | 1e-5 | 15 | 5e-6 | 15 | 5e-5 | 13 |
| $+\mathcal{D}$ | MCD | 5e-5 | 3 | 5e-5 | 12 | 5e-5 | 10 | 1e-5 | 14 | 1e-5 | 15 | 5e-5 | 13 |
| $+\mathcal{D}$ | VI | 1e-5 | 5 | 5e-6 | 13 | 5e-5 | 14 | 1e-5 | 14 | 1e-6 | 15 | 5e-5 | 5 |

Table B.9: Best hyperparameter for each model configuration and dataset pair. We denote both English and French domain-specific PLMs with $+\mathcal{D}$. The models DC, MCD, VI are from the $+\mathcal{U}$ set.

| Dataset | Sample | Classes | Label Distribution |
|---|---|---|---|
| MedABS (Schopf et al., 2023) | {text: "*Catheterization of coronary artery bypass graft from the descending aorta. The increasing frequency of reoperation for coronary artery disease has led to the use of a variety of grafts. This report describes the catheter technique for selective opacification of a saphenous vein graft from the descending thoracic aorta to the posterior coronary circulation.* ", label: "Cardiovascular diseases" } | {'Neoplasms', 'Digestive system', 'Nervous system', 'Cardiovascular', 'General pathological' } | [1925 913 1149 1804 2871] |
| MedNLI (Romanov and Shivade, 2018) | {text: "*No history of blood clots or DVTs, has never had chest pain prior to one week ago. [SEP] Patient has angina*", label: "entailment"} | {"entailment", "contradict", "neutral" } | [3744 3744 3744] |
| SMOKING (Uzuner et al., 2008) | {text: "*071962960 bh 4236518 417454 12/10/2001 12:00:00 am discharge summary unsigned dis report status : unsigned discharge summary name : sterpsap , ny unit number : 582-96-88 admission date : 12/10/2001 discharge date : 12/19/2001 principal diagnosis : prosthetic aortic valve dysfunction associated diagnoses : aortic valve insufficiency bacterial endocarditis , active principal procedure : urgent re-do aortic valve replacement and correction of left ventricular to aortic discontinuity . ( 12/13/2001 ) other procedures : aortic root aortogram ( 12/12/2001 ) cardiac ultrasound ( 12/13/2001 ) insertion dual chamber pacemaker ( 12/15/2001 ) picc line placement ( 12/18/2001 ) history and reason for hospitalization : mr. sterpsap ...*", label: "CURRENT SMOKER"} | {'CURRENT SMOKER', 'NON-SMOKER', 'PAST SMOKER', 'SMOKER', 'UNKNOWN' } | [ 27 49 24 8 190] |
| MEDMCQA (Labrak et al., 2023b) | {text: "*ans la liste suivante, quels sont les antibiotiques utilisables pour traiter une salmonellose chez un adulté?*", label: 2} | {1,2,3,4,5} | [595 528 718 296 34] |
| MORFITT (Labrak et al., 2023c) | {text: "*La survenue de complications postopératoires représente un cauchemar (bien réel), tant pour le patient que pour son chirurgien. Dès lors, quoi de plus fantasmagorique que d'administrer une « potion magique » au patient avant l'intervention pour éliminer ce risque ? Le but de cet article est de résumer l'état des connaissances actuelles concernant les bénéfices potentiels, liés à l'administration d'immunonutrition aux patients traités pour cancer urologique.....*", original_label: [ "Immunologie","Chirurgie",], label: "Immunologie"} | {'Vétérinaire', 'Étiologie', 'Psychologie', 'Chirurgie', 'Génétique', 'Physiologie', 'Pharmacologie', 'Microbiologie', 'Immunologie', 'Chimie', 'Virologie', 'Parasitologie' } | [ 82 261 32 122 40 17 152 39 242 185 104 238] |
| PxSLU (Kocabiyikoglu et al., 2022) | {text: "*antacapone 200 milligrammes 2 comprimés le matin 1 comprimé à midi 2 comprimé le soir traitement pour une durée totale de 4 semaines*", label: "medical_prescription"} | {"medical_prescription", "negate","replace", "none" } | [1276 15 82 13] |

Table B.8: Sample from each Biomedical Datasets

| | Model | | Classification | | Uncertainty | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Macro-F1(↑) | Accuracy(↑) | BS(↓) | ECE(↓) | SCE(↓) | NLL(↓) | Cov%(↑) | Entropy(↓) |
| MedABS | $-\mathcal{D}$ | DNN | 60.3633±0.003 | 60.9765±0.002 | 0.5535±0.008 | 0.1387±0.016 | 0.0683±0.004 | 1.3261±0.001 | 0.8976±0.013 | 1.5579±0.002 |
| | $-\mathcal{D}$ | DC | 60.9855±0.004 | 61.1842±0.003 | 0.5518±0.002 | 0.1342±0.007 | 0.0674±0.003 | 1.3192±0.002 | 0.9611±0.003 | 1.5556±0.001 |
| | $-\mathcal{D}$ | MCD | 60.6979±0.004 | 60.0993±0.006 | 0.5691±0.015 | 0.1503±0.014 | 0.0688±0.01 | 1.3235±0.008 | 0.9401±0.013 | 1.5542±0.002 |
| | $-\mathcal{D}$ | VI | 60.8725±0.001 | 61.1611±0.001 | 0.5531±0.006 | 0.1394±0.004 | 0.0695±0.003 | 1.3164±0.003 | 0.958±0.001 | 1.5541±0.001 |
| | $+\mathcal{D}$ | DNN | 60.8077±0.013 | 61.3343±0.01 | 0.5499±0.014 | 0.1448±0.005 | 0.0695±0.001 | 1.3201±0.014 | 0.9193±0.005 | 1.5561±0.003 |
| | $+\mathcal{D}$ | DC | 62.5642±0.009 | 62.1018±0.01 | 0.5243±0.015 | 0.1381±0.016 | 0.0624±0.007 | 1.2962±0.007 | 0.9597±0.008 | 1.5523±0.002 |
| | $+\mathcal{D}$ | MCD | 62.2038±0.022 | 62.1307±0.022 | 0.5226±0.031 | **0.1238**±0.031 | **0.0593**±0.015 | 1.3056±0.013 | **0.9666**±0.01 | 1.5562±0.002 |
| | $+\mathcal{D}$ | VI | **63.1893**±0.004 | **63.1694**±0.003 | **0.5234**±0.009 | 0.1464±0.01 | 0.0653±0.003 | **1.288**±0.006 | 0.9603±0.005 | **1.5491**±0.002 |
| MedNLI | $-\mathcal{D}$ | DNN | 73.8951±0.013 | 73.8397±0.015 | 0.3976±0.006 | 0.1278±0.02 | 0.0846±0.012 | 0.8177±0.015 | 0.9119±0.008 | 1.0156±0.008 |
| | $-\mathcal{D}$ | DC | 74.8161±0.019 | 74.8711±0.018 | 0.4242±0.021 | 0.185±0.007 | 0.1259±0.005 | 0.7945±0.014 | 0.8509±0.007 | 0.9941±0.002 |
| | $-\mathcal{D}$ | MCD | 72.8896±0.03 | 73.0192±0.03 | 0.4163±0.009 | 0.1214±0.049 | 0.0865±0.03 | 0.8298±0.037 | 0.9109±0.04 | 1.0171±0.02 |
| | $-\mathcal{D}$ | VI | 73.0816±0.022 | 73.1364±0.022 | 0.4426±0.016 | 0.185±0.023 | 0.1265±0.015 | 0.8109±0.022 | 0.857±0.035 | 0.9983±0.011 |
| | $+\mathcal{D}$ | DNN | 77.172±0.041 | 77.2386±0.039 | 0.3783±0.05 | 0.1579±0.009 | 0.107±0.007 | 0.7736±0.039 | 0.857±0.015 | 0.9952±0.008 |
| | $+\mathcal{D}$ | DC | 79.9945±0.037 | 80.0047±0.037 | **0.3375**±0.045 | 0.1392±0.005 | 0.0956±0.002 | 0.7486±0.041 | 0.8872±0.011 | 0.9924±0.011 |
| | $+\mathcal{D}$ | MCD | **80.1022**±0.014 | **80.1688**±0.014 | 0.3453±0.02 | 0.1565±0.009 | 0.1065±0.005 | **0.7437**±0.012 | 0.8654±0.004 | **0.9872**±0.001 |
| | $+\mathcal{D}$ | VI | 77.0617±0.043 | 77.1027±0.042 | 0.351±0.046 | **0.1041**±0.019 | **0.0773**±0.01 | 0.7851±0.046 | **0.9293**±0.025 | 1.0101±0.015 |
| SMOKING | $-\mathcal{D}$ | DNN | **27.1141**±0.041 | 45.8333±0.142 | 0.7724±0.054 | 0.2961±0.057 | 0.154±0.012 | 1.4298±0.106 | 0.7724±0.163 | 1.5536±0.028 |
| | $-\mathcal{D}$ | DC | 25.7924±0.041 | 46.7949±0.039 | **0.6407**±0.035 | **0.1625**±0.043 | **0.1215**±0.016 | 1.4331±0.035 | 0.9455±0.031 | 1.5791±0.01 |
| | $-\mathcal{D}$ | MCD | 26.707±0.058 | 45.8333±0.073 | 0.7609±0.077 | 0.2771±0.048 | 0.1507±0.021 | 1.4519±0.045 | 0.8942±0.058 | 1.5651±0.003 |
| | $-\mathcal{D}$ | VI | 23.4485±0.034 | 32.0513±0.043 | 0.7197±0.053 | 0.2171±0.021 | 0.15±0.023 | 1.5031±0.038 | 0.8974±0.113 | 1.5887±0.004 |
| | $+\mathcal{D}$ | DNN | 24.9822±0.041 | **51.6026**±0.071 | 0.6764±0.076 | 0.2262±0.013 | 0.1334±0.031 | 1.3928±0.068 | 0.6571±0.114 | 1.5596±0.011 |
| | $+\mathcal{D}$ | DC | 27.0293±0.033 | 47.1154±0.075 | 0.841±0.043 | 0.3441±0.053 | 0.1738±0.007 | 1.4297±0.06 | 0.7276±0.118 | 1.5419±0.02 |
| | $+\mathcal{D}$ | MCD | 25.0029±0.051 | 40.3846±0.058 | 0.6777±0.022 | 0.206±0.019 | 0.1401±0.014 | 1.482±0.014 | **0.9487**±0.04 | 1.5895±0.003 |
| | $+\mathcal{D}$ | VI | 26.1167±0.03 | 50.3205±0.094 | 0.765±0.175 | 0.3201±0.094 | 0.1584±0.045 | **1.3857**±0.094 | 0.75±0.063 | **1.5397**±0.003 |
| PxSLU | $-\mathcal{D}$ | DNN | 32.2541±0.075 | 88.2452±0.012 | 0.5743±0.077 | 0.4556±0.094 | 0.2955±0.014 | 1.2807±0.05 | 0.995±0.004 | 1.3821±0.003 |
| | $-\mathcal{D}$ | DC | 34.1464±0.026 | 84.2989±0.05 | 0.4599±0.088 | 0.3936±0.047 | 0.2354±0.03 | 1.2154±0.062 | **1.0**±0.0 | 1.3768±0.007 |
| | $-\mathcal{D}$ | MCD | 33.211±0.067 | 88.6902±0.018 | 0.5232±0.103 | 0.4852±0.079 | 0.2615±0.027 | 1.2571±0.062 | **1.0**±0.0 | 1.3806±0.004 |
| | $-\mathcal{D}$ | VI | 25.9883±0.041 | 88.9169±0.013 | 0.5393±0.021 | 0.5014±0.026 | 0.2552±0.007 | 1.2666±0.014 | **1.0**±0.0 | 1.3814±0.001 |
| | $+\mathcal{D}$ | DNN | 33.1131±0.097 | 80.1763±0.238 | 0.5389±0.116 | 0.3929±0.057 | 0.2867±0.037 | 1.2548±0.06 | 0.9831±0.018 | 1.38±0.003 |
| | $+\mathcal{D}$ | DC | 40.3372±0.07 | 89.1184±0.039 | 0.2649±0.127 | 0.2576±0.105 | 0.1568±0.058 | 1.0539±0.111 | 0.9997±0.001 | 1.3496±0.021 |
| | $+\mathcal{D}$ | MCD | 34.1571±0.029 | 89.1436±0.026 | 0.5403±0.043 | 0.5074±0.015 | 0.2663±0.013 | 1.2694±0.026 | **1.0**±0.0 | 1.3821±0.002 |
| | $+\mathcal{D}$ | VI | **41.8279**±0.073 | **91.0999**±0.015 | **0.1634**±0.051 | **0.1403**±0.064 | **0.0861**±0.029 | **0.9464**±0.066 | 0.9958±0.004 | **1.3246**±0.019 |
| MEDMCQA | $-\mathcal{D}$ | DNN | 28.5727±0.03 | 63.88±0.055 | 0.6787±0.1 | 0.3256±0.043 | 0.1575±0.021 | 1.5347±0.062 | 0.9625±0.033 | 1.6063±0.003 |
| | $-\mathcal{D}$ | DC | **32.0291**±0.003 | 63.5584±0.007 | **0.4822**±0.015 | **0.165**±0.01 | **0.1099**±0.0 | **1.3846**±0.009 | 0.9764±0.007 | **1.5888**±0.001 |
| | $-\mathcal{D}$ | MCD | 28.3648±0.029 | 61.3612±0.103 | 0.7533±0.044 | 0.3819±0.084 | 0.1518±0.02 | 1.5848±0.024 | **1.0**±0.0 | 1.6091±0.0 |
| | $-\mathcal{D}$ | VI | 23.1977±0.042 | 48.5531±0.046 | 0.7499±0.023 | 0.242±0.033 | 0.1329±0.004 | 1.5822±0.013 | **1.0**±0.0 | 1.6089±0.0 |
| | $+\mathcal{D}$ | DNN | 28.1549±0.045 | 61.0932±0.089 | 0.6859±0.12 | 0.3026±0.009 | 0.1582±0.01 | 1.5388±0.077 | 0.9775±0.02 | 1.6064±0.004 |
| | $+\mathcal{D}$ | DC | 29.7558±0.07 | 60.343±0.103 | 0.6687±0.17 | 0.2973±0.069 | 0.1278±0.018 | 1.5216±0.122 | 0.9893±0.019 | 1.6025±0.012 |
| | $+\mathcal{D}$ | MCD | 31.0912±0.016 | **68.4352**±0.033 | 0.5541±0.115 | 0.3122±0.059 | 0.1477±0.031 | 1.4543±0.081 | 0.9936±0.011 | 1.5999±0.007 |
| | $+\mathcal{D}$ | VI | 23.1243±0.035 | 49.8553±0.031 | 0.7415±0.017 | 0.2336±0.026 | 0.1222±0.008 | 1.5765±0.01 | **1.0**±0.0 | 1.6085±0.0 |
| MORFITT | $-\mathcal{D}$ | DNN | 49.7506±0.009 | 59.038±0.012 | 0.6499±0.022 | 0.2323±0.021 | 0.0398±0.005 | 2.0748±0.015 | 0.796±0.045 | 2.4454±0.003 |
| | $-\mathcal{D}$ | DC | 55.4551±0.01 | 62.5306±0.008 | 0.6134±0.003 | 0.2243±0.003 | 0.0425±0.001 | 2.0332±0.006 | 0.8775±0.014 | 2.4411±0.001 |
| | $-\mathcal{D}$ | MCD | 48.3269±0.008 | 57.3529±0.008 | 0.6309±0.021 | 0.1519±0.05 | 0.0464±0.007 | 2.2692±0.03 | **0.9856**±0.006 | 2.4767±0.003 |
| | $-\mathcal{D}$ | VI | 53.0834±0.014 | 61.6728±0.01 | 0.6408±0.042 | 0.2571±0.039 | 0.0477±0.006 | **2.0245**±0.007 | 0.7724±0.047 | **2.4369**±0.004 |
| | $+\mathcal{D}$ | DNN | 53.4963±0.019 | 61.8015±0.014 | 0.6081±0.017 | 0.2098±0.014 | 0.0363±0.002 | 2.0538±0.015 | 0.8334±0.01 | 2.4453±0.002 |
| | $+\mathcal{D}$ | DC | **56.4418**±0.018 | **62.9596**±0.02 | 0.6148±0.027 | 0.2325±0.018 | 0.0433±0.003 | 2.0251±0.015 | 0.8667±0.03 | 2.4394±0.001 |
| | $+\mathcal{D}$ | MCD | 51.8519±0.015 | 60.5392±0.006 | 0.5718±0.003 | 0.0687±0.022 | 0.0298±0.0 | 2.1426±0.01 | 0.9651±0.005 | 2.4629±0.002 |
| | $+\mathcal{D}$ | VI | 54.2993±0.011 | 62.7145±0.01 | **0.5346**±0.008 | **0.0488**±0.018 | **0.0279**±0.002 | 2.1064±0.014 | 0.9752±0.007 | 2.4602±0.002 |

Table B.10: Comparison for text classification performance and uncertainty-awareness. We report the mean of 10 seed runs for all the metrics. We denote best score with **bold** and second best with underline. We denote both English and French domain-specific PLMs with $+\mathcal{D}$. The models DC, MCD, VI are from the $+\mathcal{U}$ set.

## Interpretation of Medical Models

**Models Description** Throughout the study, we utilize the collection of miniature distilled BERT models. Turc et al. (2019) proposed 24 models, each of which differ in terms of the numerical count in the architecture of BERT-base-case model Devlin et al. (2019). The number of parameters for each of models can be seen in Table B.11.

**Reproducibility details.** All the datasets are trained with default setting of the Turc et al. (2019) models for 20 epochs (except for SNLI which is trained only for 2 epochs due to large dataset size.

| | $N_{\text{param}}$ | $\theta_{\text{i}}$ | | $N_{\text{param}}$ | $\theta_{\text{i}}$ |
|---|---|---|---|---|---|
| BU_L-2_H-128_A-2 | 4385920 | $m_1$ | BU_L-2_H-512_A-8 | 22458880 | $m_{13}$ |
| BU_L-4_H-128_A-2 | 4782464 | $m_2$ | BU_L-4_H-512_A-8 | 28763648 | $m_{14}$ |
| BU_L-6_H-128_A-2 | 5179008 | $m_3$ | BU_L-6_H-512_A-8 | 35068416 | $m_{15}$ |
| BU_L-8_H-128_A-2 | 5575552 | $m_4$ | BU_L-2_H-768_A-12 | 38603520 | $m_{16}$ |
| BU_L-10_H-128_A-2 | 5972096 | $m_5$ | BU_L-8_H-512_A-8 | 41373184 | $m_{17}$ |
| BU_L-12_H-128_A-2 | 6368640 | $m_6$ | BU_L-10_H-512_A-8 | 47677952 | $m_{18}$ |
| BU_L-2_H-256_A-4 | 9591040 | $m_7$ | BU_L-4_H-768_A-12 | 52779264 | $m_{19}$ |
| BU_L-4_H-256_A-4 | 11170560 | $m_8$ | BU_L-12_H-512_A-8 | 53982720 | $m_{20}$ |
| BU_L-6_H-256_A-4 | 12750080 | $m_9$ | BU_L-6_H-768_A-12 | 66955008 | $m_{21}$ |
| BU_L-8_H-256_A-4 | 14329600 | $m_{10}$ | BU_L-8_H-768_A-12 | 81130752 | $m_{22}$ |
| BU_L-10_H-256_A-4 | 15909120 | $m_{11}$ | BU_L-10_H-768_A-12 | 95306496 | $m_{23}$ |
| BU_L-12_H-256_A-4 | 17488640 | $m_{12}$ | BU_L-12_H-768_A-12 | 109482240 | $m_{24}$ |

Table B.11: Model complexity as measured by $N_{\text{param}}$ (number of parameters).

## Concept Alignment Gap

For pretrained language models, we embedded the given clinical abstract and the given tier list of hallmark labels using the embedding model. We then use utilize the similarity metrics from `sentence-transformer` library to compute the pair wise score for each hallmark label with the clinical abstract. We select the hallmark label which obtains the highest score from the list of labels.

For large language models, we obtain the predictions using zero shot prompting. For a provided tier list of HoI labels, we prompt the LLM with the prompt template shown in fig. 5.18. Further, we preprocess the answers obtained from the LLM models via GPT-3.5 turbo to obtain the prediction from the raw output generated by the model. Table B.12 lists the huggingface model id of all the models used in the experimentation to enable better reproducibility.

| | | **huggingface model_id** |
|---|---|---|
| PLMs | ClinicalBigBird | `yikuan8/Clinical-BigBird` |
| | PubmedBERT | `NeuML/pubmedbert-base-embeddings` |
| | BioBERT | `dmis-lab/biobert-v1.1` |
| | ClinicalBERT | `emilyalsentzer/Bio_ClinicalBERT` |
| | SciBERT | `allenai/scibert_scivocab_uncased` |
| LLMs | Llama-3-8B-It | `meta-llama/Meta-Llama-3-8B-Instruct` |
| | Gemma-2-9B | `google/gemma-2-9b-it` |
| | Med-Qwen-7B | `Echelon-AI/Med-Qwen2-7B` |
| | DeepSeek-R1 | `deepseek-ai/DeepSeek-R1-Distill-Llama-8B` |
| | BioMistral-7B | `ZiweiChen/BioMistral-Clinical-7B` |

Table B.12: Model Huggingface ids used in the experimentation

## Difficulty Perception Gap

**Implementation details** Our use of all preexisting research artifacts is consistent with their corresponding licenses. We trust creators of said artifacts to have handled any personally identifying information that the artifacts may contain.

**Data** As noted above, we use SNLI (Bowman et al. (2015): retrieved from HuggingFace), MNLI (Williams et al. (2018a), retrieved from HuggingFace) and ChaosNLI (Nie et al. (2020b); retrieved from GitHub). We remove items without public labels from SNLI and MNLI, as well as datapoints with no majority label from DynaSent.

**Models** All models are implemented with HuggingFace (HF; Wolf et al. (2020); Lhoest et al. (2021)). As per default HF implementations, for the 1B pool of models, classifiers rely on the last token in the input; for the <1B model pool, we use the first token. All experiments are supervised full fine-tuning processes using learned linear projections as classification heads. Models are trained on a V100 NVIDIA GPU, for an individual runtime of ≤15 hours for any individual model.

| **<1B SNLI Models** | |
|---|---|
| Number of epochs | 2 |
| Batch size | 16 |
| **1B SNLI Models** | |
| Number of epochs | 10 |
| Batch size | 16 |
| **All MNLI Models** | |
| Number of epochs | 5 |
| Batch size | 1 |
| Gradient accumulation | 16 |
| Warmup ratio | 0.1 |
| Learning rate | 1e-6 |

Table B.13: Hyperparameters used for all SNLI and MNLI models, grouped by model size and dataset.

Hyperparameters are listed in Table B.13. Any hyperparameter not listed in Table B.13 was left to its default value as listed in the HF documentation.

## B.3 Code Repositories

| Section | Dataset | Repository |
|---|---|---|
| Section 3.2.1 | SocialDisNER | https://github.com/amansinha09/social-ner |
| Section 3.2.2 | CMED | https://github.com/amansinha09/nncc |
| Section 3.2.3 | REFOMED | https://github.com/ATILF-UMR7118/pRAGe |
| Section 3.3.1 | CMED | https://github.com/amansinha09/nncc |
| Section 3.3.2 | CMED | https://github.com/amansinha09/nncc |
| Section 3.4.1 | PUBMED | |
| Section 3.4.2 | INSOMNIA | https://github.com/amansinha09/SMM4H-2025 |
| Section 3.4.2 | RADNLP | https://github.com/amansinha09/RADREP |
| Section 3.5.1 | MIMIC-III, PhysioNet 2012 | https://github.com/Rohit102497/SLAN_Previous |
| Section 4.2.1 | SocialDisNER | https://github.com/amansinha09/social-ner |
| Section 4.2.2 | REFOMED | https://github.com/ATILF-UMR7118/pRAGe |
| Section 4.2.3 | PUBMED | |
| Section 4.3.1 | MIMIC-III | https://github.com/amansinha09/mSLAN |
| Section 4.4.1 | CMED | https://github.com/amansinha09/nncc |
| Section 5.2.1 | MedABS, MedNLI, SMOKING, MEDMCQA, MORFIT, PxSLU | https://github.com/amansinha09/ublue |
| Section 5.2.2 | SST2, ChaosNLI, HateXplain, SNLI, MNLI, MedABS, MedNLI | https://github.com/amansinha09/cerberus |
| Section 5.3.1 | HoI | https://github.com/ICANS-Strasbourg/immuNLP |
| Section 5.3.2 | SNLI, MNLI, ChaosNLI | https://github.com/Helsinki-NLP/data-cplx-unc |

Table B.14: Datasets and code repositories utilized in this thesis.