Enhancing LLMs for Identifying and Prioritizing Important Medical Jargons from Electronic Health Record Notes Utilizing Data Augmentation

Won Seok Jang, Zonghai Yao, Sharmin Sultana, Hieu Tran, Zhichao Yang, Sunjae Kwon, Hong Yu

Submitted to: JMIR Preprints
on: March 07, 2025

Table of Contents

Enhancing LLMs for Identifying and Prioritizing Important Medical Jargons from Electronic Health Record Notes Utilizing Data Augmentation

Won Seok Jang[1*] MSc; Zonghai Yao[2*] MSc; Sharmin Sultana[1*] BSc; Hieu Tran[2] MSc; Zhichao Yang[2] PhD; Sunjae Kwon[2] MSc; Hong Yu[1, 3] PhD

[1] Miner School of Computer and Information Sciences University of Massachusetts Lowell Lowell US
[2] Manning College of Information & Computer Sciences University of Massachusetts Amherst Amherst Center US
[3] Center for Healthcare Organization and Implementation Research Edith Nourse Rogers Memorial Veterans Hospital Bedford US
[*]these authors contributed equally

Corresponding Author:
Won Seok Jang MSc

Miner School of Computer and Information Sciences
University of Massachusetts Lowell
University of Massachusetts Lowell Richard A. Miner School of Computer & Information Sciences Dandeneau Hall 1 University Avenue
Lowell
US

Abstract

Background: OpenNotes allows patients to access their electronic health record (EHR) notes through online patient portals. However, EHR notes contain abundant medical jargon, which can be difficult for patients to comprehend. One way to improve comprehension is by reducing information overload and helping patients focus on the medical terms that matter most to them.

Objective: In this study, we evaluated both closed-source and open-source Large Language Models (LLMs) for extracting and prioritizing medical jargon from EHR notes relevant to individual patients, leveraging prompting techniques, fine-tuning, and data augmentation.

Methods: We evaluated the performance of closed-source and open-source LLMs on a dataset of 106 expert-annotated EHR notes. We tested various combinations of settings, including: i) general and structured prompts, ii) zero-shot and few-shot prompting, iii) fine-tuning, and iv) data augmentation. To enhance the extraction and prioritization capabilities of open-source models in low-resource settings, we applied data augmentation using ChatGPT and integrated a ranking technique to refine the training process. Additionally, to measure the impact of dataset size, we fine-tuned the models by incrementally increasing the size of the augmented dataset from 10 to 10,000 and tested their performance. The effectiveness of the models was assessed using 5-fold cross-validation, providing a comprehensive evaluation across various settings. We report the F1 score and Mean Reciprocal Rank (MRR) for performance evaluation.

Results: Among the compared strategies, fine-tuning and data augmentation generally demonstrated higher performance than other approaches. Although the highest F1 score of 0.433 was achieved by GPT-4 Turbo, the highest MRR score of 0.746 was observed with Mistral7B when data augmentation was applied. Notably, by using f ine-tuning or data augmentation, open-source models were able to outperform closed-source models. Additionally, achieving the highest F1 score did not always correspond to the highest MRR score. We analyzed our experiment from several perspectives. First, few-shot prompting showed an advantage over zero-shot prompting in vanilla models. Second, when comparing general and structured prompts, each model exhibited different preferences. Third, finetuning improved zero-shot performance but sometimes degraded few-shot performance. Lastly, data augmentation yielded performance comparable to or even surpassing that of other strategies.

Conclusions: The evaluation of both closed-source and open-source LLMs highlighted the effectiveness of prompting strategies, fine-tuning, and data augmentation in enhancing model performance in low-resource scenarios.

(JMIR Preprints 07/03/2025:73449)
DOI: https://doi.org/10.2196/preprints.73449

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

Enhancing LLMs for Identifying and Prioritizing Important Medical Jargons from Electronic Health Record Notes Utilizing Data Augmentation

Won Seok Jang*, MSc[1], Sharmin Sultana*, BSc[1], Zonghai Yao*, MSc[2] Hieu Tran, BSc[2], Zhichao Yang, PhD[2], Sunjae Kwon, MSc[2], Hong Yu, PhD[1,3]

[1] Miner School of Computer and Information Sciences, UMass Lowell, MA, USA

[2] Manning College of Information and Computer Sciences, UMass Amherst, MA, USA

[3] Center for Healthcare Organization and Implementation Research, VA Bedford Health Care, MA, USA

These authors contributed equally *: Won Seok Jang, Sharmin Sultana and Zonghai Yao

Corresponding author: Hong Yu, Ph.D.

Phone: 1 978-934-3620 Email: Hong_Yu@uml.edu

Abstract

Background: OpenNotes allows patients to access their electronic health record (EHR) notes through online patient portals. However, EHR notes contain abundant medical jargon, which can be difficult for patients to comprehend. One way to improve comprehension is by reducing information overload and helping patients focus on the medical terms that matter most to them.

Objective: In this study, we evaluated both closed-source and open-source Large Language Models (LLMs) for extracting and prioritizing medical jargon from EHR notes relevant to individual patients, leveraging prompting techniques, fine-tuning, and data augmentation.

Materials and Methods: We evaluated the performance of closed-source and open-source LLMs on a dataset of 106 expert-annotated EHR notes. We tested various combinations of settings, including: i) general and structured prompts, ii) zero-shot and few-shot prompting, iii) fine-tuning, and iv) data augmentation. To enhance the extraction and prioritization capabilities of open-source models in low-resource settings, we applied data augmentation using ChatGPT and integrated a ranking technique to refine the training process. Additionally, to measure the impact of dataset size, we fine-tuned the models by incrementally increasing the size of the augmented dataset from 10 to 10,000 and tested their performance. The effectiveness of the models was assessed using 5-fold cross-validation, providing a comprehensive evaluation across various settings. We report the F1 score and Mean Reciprocal Rank (MRR) for performance evaluation.

Results and Discussions: Among the compared strategies, fine-tuning and data augmentation generally demonstrated higher performance than other approaches. Although the highest F1 score of 0.433 was achieved by GPT-4 Turbo, the highest MRR score of 0.746 was observed with Mistral7B when data augmentation was applied. Notably, by using fine-tuning or data augmentation, open-source models were able to outperform closed-source models. Additionally, achieving the highest F1 score did not always correspond to the highest MRR score. We analyzed our experiment from several perspectives. First, few-shot prompting showed an advantage over zero-shot prompting in vanilla models. Second, when comparing general and structured prompts, each model exhibited different preferences. Third, finetuning improved zero-shot performance but sometimes degraded few-shot performance. Lastly, data augmentation yielded performance comparable to or even surpassing that of other strategies.

Conclusion: The evaluation of both closed-source and open-source LLMs highlighted the effectiveness of prompting strategies, fine-tuning, and data augmentation in enhancing model

performance in low-resource scenarios. Keywords: LLMs, Data Augmentation, EHR Comprehension, Patient Education, Patient Engagement

Word count: 3813

Introduction

Electronic Health Record (EHR) notes serve as valuable sources of information that can significantly benefit patients. Programs like OpenNotes [1] and the Blue Button [2] initiative empower patients by providing access to their EHR notes [3, 4, 5, 6, 7, 8, 9]. Nevertheless, the benefits of accessing EHR notes can diminish significantly if patients do not comprehend their content [10, 11, 12, 13, 14, 15]. EHR notes are lengthy and filled with medical jargon [16, 17, 18, 19], which can be difficult to comprehend for the average U.S. adult, whose reading ability is around the 7th to 8th-grade level [20, 21, 22, 23, 24]. Therefore, supportive technologies are needed to assist patients in understanding EHR content [25, 26], focusing on linking medical terms to lay-friendly terms [27, 28, 29], consumer-oriented definitions [18], and educational materials [30]. Early studies have demonstrated that such interventions significantly enhance patient understanding [27, 18]. However, initial methods primarily relied on frequency- and context-based approaches to identify unfamiliar terms and propose simpler synonyms [27, 28, 29]. Identifying and extracting complex medical jargon from EHR notes is a crucial step toward improving patients' comprehension, ultimately enhancing patient engagement and reducing anxiety about their health [18, 31, 32, 33].

Notably, not all medical jargon extracted from EHR notes holds equal clinical importance [34, 35]. Existing tools, such as MetaMap [36], ScispaCy [37], medspaCy [38], and QuickUMLS [39], are effective at extracting medical terms—typically predefined terms from the Unified Medical Language System (UMLS)[40]—but often fail to prioritize these terms based on their relevance to individual patients, treating all terms as equally important[27, 28, 29]. In previous work, we asked physicians to identify medical jargon terms from EHR notes that are important to patients [34]. Our results showed that physicians were able to consistently identify 5 to 10 medical jargon terms from each EHR note and rank each term based on its importance to patients [34]. Furthermore, we developed feature-rich traditional machine learning models (e.g., support vector machines) to identify such terms [34]. However, our previous work did not focus on ranking jargon terms based on their importance to individual patients within an EHR note.

In this study, we propose large language model (LLM)-based NLP approaches to identify and rank jargon terms from EHR notes based on their importance to patients. LLMs have demonstrated tremendous promise in biomedical NLP applications [41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53] due to their exceptional generalizability and performance. However, applications in medical term extraction have primarily focused on tasks such as biomedical named entity recognition (BioNER) [54, 55, 56], rather than on prioritizing terms most relevant to patients, which is crucial for enhancing communication between patients and healthcare providers.

The key contributions of this paper are as follows:

- To the best of our knowledge, we are the first to conduct a comprehensive evaluation of both closed-source and open-source LLMs to assess their effectiveness in identifying medical jargon from EHR notes that are important for patients.
- We developed a novel data augmentation technique by leveraging MIMIC discharge summaries to address the challenges of training in low-resource settings, resulting in significant performance improvements. Notably, our method enabled smaller open-source LLMs (<10B parameters) to outperform much larger models from the GPT and Claude3

families.

- In the discussion section, we provide an in-depth analysis of the results from both quantitative and qualitative perspectives, focusing on common strategies for improving LLM performance, such as zero-shot and few-shot learning, prompt engineering, scaling laws, domain-adaptive training, and data augmentation, and their impact on this specific task.

Related Work

Identifying jargon terms important to patients is part of the biomedical Named Entity Recognition (BioNER) task, which involves identifying predefined entities in a text and labeling each token with the corresponding entity. Medical entities encompass categories such as diseases, medications, treatments, lab tests, and more [57, 58]. Studies such as [59, 60, 61, 62] have introduced language models for BioNER tasks, while more recent studies [54, 55, 56, 63, 64] have explored the application of large language models (LLMs) in BioNER. However, BioNER tasks primarily focus on extracting entities without considering their importance and relevance to the personal needs of patients, which distinguishes them from our objective.

MedJex [32] fine-tunes pre-trained language models (PLMs), such as BERT [65], RoBERTa [60], BioClinicalBERT [66], and BioBERT [59], on a domain-specific corpus. It leverages Wikipedia hyperlink spans during pretraining and transfers the learned weights to a target model fine-tuned on MedJ, an expert-annotated medical dataset. More recent studies [67] have investigated whether large language models (LLMs), such as ChatGPT [68], can outperform baseline PLMs (e.g., MedJex [32] and SciSpacy [37]) in extracting personalized medical jargon. Similarly, GAMedX [69], a medical data extractor utilizing LLMs (Mistral 7B and Gemma 7B), employs chained prompts to navigate the complexities of specialized medical jargon. Other works [70, 71, 72] have demonstrated how LLMs can enhance the readability of EHR notes by extracting medical jargon.

This work also shares similarities with topic modeling, a task that extracts topics from input text. Using unsupervised learning algorithms, topic modeling can identify both explicit and implicit themes within a text corpus [73, 74, 75]. Through topic modeling, a text can be represented by multiple keywords or topics, which can then be incorporated into supervised models. However, topic modeling heavily relies on term frequencies and may easily overlook important terms that are clinically relevant to individual patients.

Among the most relevant works, such as FOCUS [34], ADS [76], and FIT [35], FOCUS [34] employs MetaMap [77] to extract medical jargon from EHR notes and utilizes feature-rich learning-to-rank techniques to determine whether the terms are important. However, none of the previous works have identified and ranked medical jargon terms in a note-specific manner. This is an important task, as ranking terms based on their relevance to a specific note may help the patient comprehend the note by linking important jargon terms to their lay definitions [78], or help generate patient-friendly after-visit summaries [79].

Materials and Methods

Overview

We evaluated both closed-source and open-source LLMs for their efficacy in extracting key information from annotated medical notes, aiming to assess performance across different strategies. Figure 1 provides an overview of our experiments, which leverage physician-annotated medical notes, closed- and open-source LLMs, and In-Context Learning (ICL). We examined the effects of various prompting styles, fine-tuning, and data augmentation to enhance model performance.

Data source

Our gold-standard dataset consists of 106 medical notes, each annotated by two physicians [34, 35]. This EHR note dataset comprises text reports across six medical categories: Cancer, COPD, Diabetes, Heart Failure, Hypertension,

Main Diagnosis	Note Counts	Median Jargon Counts
Cancer	20	14
COPD	19	8
Diabetes	19	9
Hypertension	21	7
Liver Failure	15	10
Heart Failure	12	9

Table 1: Gold-Standard Dataset Description. The dataset consists of 106 notes from patients diagnosed with Cancer, Chronic Pulmonary Obstructive Disease (COPD), Diabetes, Hypertension, Liver Failure and Heart Failure. The median jargon counts for each categories are illustrated.

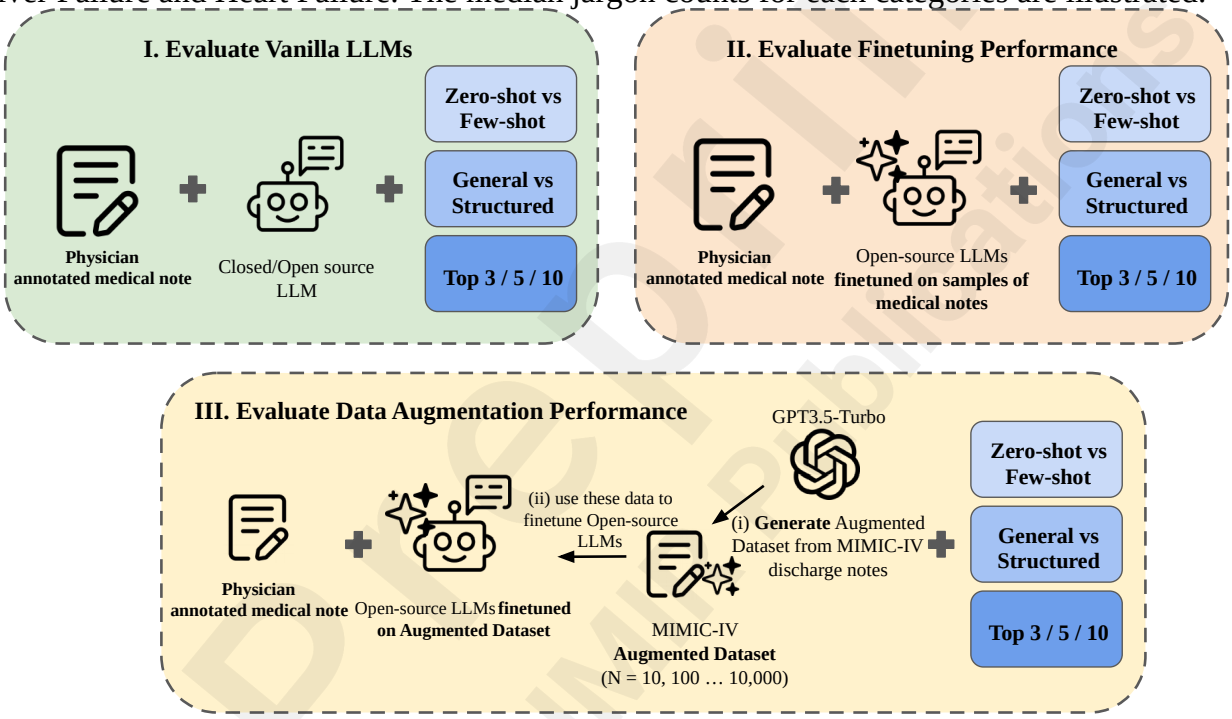

Figure 1: The evaluation workflow for closed and open-Source LLMs. We evaluate the performance of the LLMs in three distinctive settings. I. We assess the performance of closed- and open-source models by varying prompts and extraction tasks. II. Next, we fine-tune the open-source models using the same variations. III. Finally, we apply data augmentation to fine-tune the open-source LLMs and evaluate them under the same varying settings. Performance is measured using F1 and MRR scores through 5-fold cross-validation.

and Liver Failure (Table 1). Each medical note includes detailed patient information and is accompanied by physician annotations highlighting the most critical terms or phrases relevant to the patient's health status. Figure 2 presents a snippet of a sample EHR note from the gold-standard dataset.

Experimental Setup

Closed-Source and Open-Source LLMs We used both publicly available LLMs (open-source

LLMs) and proprietary models that are not publicly available (closed-source LLMs). The open-source LLMs included Mistral7B [80], BioMistral7B [81], and Llama 3.1 8B [82]. For proprietary models, we utilized six closed-source LLMs, three of which were from OpenAI [68]: GPT-4 Turbo [83], GPT-3.5 Turbo [84], and GPT-4o Mini [85]. Additionally, we conducted experiments using the Claude 3.0 models (Sonnet, Opus, and Haiku) developed by Anthropic [86].

Zero-shot vs Few-Shot We used both zero-shot and few-shot prompts, also known as ICL, to compare the performance of the LLMs. For zero-shot prompting, we provided the model with general instructions, whereas for few-shot prompting, we included two examples randomly selected from the gold-standard note dataset.

General and Structured Prompts To evaluate the model's performance, we implemented two distinct types of prompts: general prompts and structured prompts, each designed to assess how effectively the model could extract relevant clinical information from medical notes. For few-shot prompting, we included two annotated examples to guide the model in extracting and ranking terms effectively, whereas zero-shot prompts relied solely on instructions without prior examples. Figure 6 presents an example of a structured prompt. In this approach, the prompts are explicitly designed to closely align with the original task defined in the gold-standard dataset. The structured prompts instruct the LLMs to extract key medical conditions or diagnoses, followed by the relevant medications associated

An Excerpt from Annotated EHR Note

Ms. XXX returns for follow-up of her rheumatoid arthritis [1] and a history of B-cell lymphoma. Since her last visit, Ms. XXX continues to have variable days of pain related to her postherpetic neuralgia [4]. She is frustrated by the persistent pain and wonders whether any of her medications for this is very helpful. She is interested in slowly tapering off her narcotics. As for her rheumatoid arthritis, she notes increasing symptoms in her hands and most recently, her knees and feet. She wonders if she should receive another dose of rituximab [2.1]. She just saw Dr. Y for follow-up of her large cell lymphoma [5]. She mentions that her labs revealed increased serum calcium, and she was told to hold her calcium supplements and vitamin D. As for her osteoporosis [3], she had an appointment with Dr. Z recently but then decided to get denosumab [2.2] here in our clinic. She is still concerned about the potential side effects of the medication.

Ms. XXX's rheumatoid arthritis [1] symptoms have recently increased, particularly with the development of with synovitis [2], particularly in her left knee. The timing from her last dose of rituximab [2.1] is almost exactly a year. She and Dr. agree that it is reasonable to consider another infusion of this medication since she tolerated it well. The protocol would be for her RA rather than for her lymphoma. She is interested in receiving this through Dr. YY's office.

Figure 2: A sample EHR note where physicians identified important medical terms. Diagnoses/conditions are highlighted in yellow, while medications, tests and procedures associated with those diagnoses are marked in green, accompanied by their respective rankings with those conditions, ensuring that the model replicates the annotation style of the gold-standard data.

In contrast, general prompts provide a more flexible and broader context, as illustrated in Figure 7. These prompts instruct the model to extract key medical terms without explicitly differentiating between conditions and medications, assigning the same base rank to both. This approach allows the LLM to interpret the extraction task more broadly, offering insights into how well the model generalizes its understanding of medical terminology when provided with less specific guidance.

Fine-tuning with Low-Rank Adaptation (LoRA) To improve the performance of open-source LLMs (Mistral7B, Biomistral7B, and Llama 3.1 8B), we conducted LoRA [87] based parameter-efficient finetuning (PEFT). LoRA is an efficient fine-tuning technique that allows models to be adapted to specific tasks without the need to update all of the model's parameters. Instead, it applies low-rank updates to specific layers, reducing the computational cost and memory usage typically

associated with traditional fine-tuning methods. The training was done with a batch size of 1 per device and gradient accumulation over 128 steps, and low-rank dimension was set to 64. The learning rate was configured at $3e-4$, and the model was trained over 100 epochs to allow the models to converge effectively on the task-specific patterns present in the dataset.

Data Augmentation with MIMIC-IV Discharge Notes Annotation by domain experts is expensive, and data augmentation using AI-generated data can help alleviate this challenge [88]. ICL, or few-shot prompting, integrates task examples directly into input prompts, allowing models to observe patterns and generalize from limited examples to effectively handle new, unseen data [89]. We created an augmented dataset using the ICL technique, which was then used to refine the models (Supplementary Figure 8). This augmented dataset was derived from discharge notes in the MIMIC-IV clinical database [90]. A subset of MIMIC discharge notes was randomly selected, and ChatGPT (GPT-3.5 Turbo) [68] was used to process and rank key terms based on their importance for patient understanding. This extraction process was guided by the ICL framework, where two annotated notes from our gold-standard dataset were provided as examples to instruct the model on identifying and prioritizing terms. These examples ensured that the extracted terms were not only accurate but also contextually significant, reflecting the kind of information most beneficial for patient comprehension. The resulting augmented dataset provided a broader training base for fine-tuning the open-source LLMs.

Determining the size of the augmented dataset To explore the effects of augmented dataset size, we progressively increased the dataset across four scales: 10, 100, 1,000, and 10,000 notes. By leveraging the original gold-standard annotations along with AI-generated augmented data, we conducted a comprehensive evaluation of model performance across varying data sizes. Our objective was to enhance the robustness of LLMs in extracting critical medical information from EHRs to improve patient care.

Baseline Models We evaluated the performance of the open-source LLMs against several prominent baselines in the field of generative language models. GPT-2 [91], a 1.5 billion-parameter unsupervised transformer model trained on 8 million unstructured data samples, is capable of performing various tasks without task-specific training. BioGPT [92], a domain-specific GPT-2-based model trained on biomedical research articles, excels in biomedical question answering, data extraction, and text generation, demonstrating versatility in domain-specific downstream tasks.

Evaluation Metrics We used Precision (1) , Recall (2), F1 (3) and Mean Reciprocal Rank (MRR) (4). We conduct 5-fold cross-validation to measure the performance and report the average score with the confidence intervals provided in the Appendix section. Precision and recall gave insights into the model's ability to cover the annotated terms. The macro F1 score allowed us to evaluate the model's balanced performance, whereas the MRR metric, in particular, evaluated how well the model ranked the terms according to their relevance, reflecting the model's alignment with the annotated order of terms. We also used relaxed string matching to determine the true labels because exact matching strictly requires an exact string correspondence between extracted key phrases and the gold standard, which can significantly underestimate performance. [93, 94].

$$Precision = \frac{TP}{TP + FP} \quad (1) \qquad Recall = \frac{TP}{TP + FN} \quad (2)$$

$$F1 = 2 \times \frac{(P \cdot R)}{(P + R)} \quad (3)$$

$$MRR = \frac{1}{¿Q∨¿ \sum_{i=1}^{¿Q∨¿ \frac{1}{rank_i}}} \quad (4)¿¿$$

Hardware Settings All experiments were performed with two Nvidia A100 GPUs, each with 40 GB of memory, an Intel Xeon Gold 6230 CPU, and 192 GB of RAM.

Experimental Results

Table 2 presents the results from both closed- and open-source models, with confidence intervals provided in Supplementary Table 5. The highest F1 score, 0.433 (CI 95The highest MRR, 0.746 (CI 95For both F1 and MRR, the highest scores were achieved using few-shot prompts. Notably, in vanilla models—whether closed- or open-source—few-shot prompts consistently outperformed zero-shot prompts in terms of F1, regardless of the top-N extraction setting. Most of our results were statistically significant ($p < 0.05$) (Supplementary Table 5); however, this trend was weaker for MRR. Interestingly, in many settings, the highest scores were obtained from open-source models. Furthermore, with fine-tuning or data augmentation, some open-source models achieved higher F1 scores than closed-source models in zero-shot prompts.

Models	Prompt	Zero-Shot						Few-Shot					
		top3		top5		top10		top3		top5		top10	
		F1	MRR	F1	MRR	F1	MRR	F1	MRR	F1	MRR	F1	MRR
Close-Source LLMs													
GPT-4 Turbo	general	0.306	0.562	0.359	0.601	0.433	0.640	0.324	0.600	0.368	0.626	0.433	0.621
	structured	0.317	0.643	0.347	0.668	0.370	0.691	0.362	0.673	0.398	0.705	0.424	0.691
GPT-4o Mini	general	0.320	0.617	0.355	0.667	0.392	0.691	0.329	0.664	0.369	0.700	0.383	0.708
	structured	0.303	0.620	0.329	0.617	0.335	0.610	0.328	0.666	0.363	0.668	0.381	0.684
GPT-3.5 Turbo	general	0.261	0.478	0.308	0.582	0.324	0.602	0.343	0.649	0.359	0.655	0.378	0.669
	structured	0.303	0.575	0.329	0.573	0.343	0.572	0.354	0.661	0.367	0.633	0.398	0.661
Claude 3.0 Sonnet	general	0.289	0.554	0.328	0.556	0.355	0.557	0.308	0.596	0.345	0.598	0.399	0.639
	structured	0.304	0.666	0.317	0.673	0.323	0.668	0.354	0.683	0.357	0.684	0.371	0.692
Claude 3.0 Haiku	general	0.284	0.613	0.320	0.629	0.345	0.620	0.320	0.672	0.358	0.690	0.387	0.683
	structured	0.248	0.584	0.271	0.591	0.281	0.614	0.357	0.668	0.364	0.715	0.349	0.693
Claude 3.0 Opus	general	0.340	0.707	0.365	0.698	0.390	0.713	0.316	0.615	0.359	0.661	0.412	0.709
	structured	0.289	0.617	0.344	0.643	0.387	0.664	0.319	0.646	0.331	0.551	0.387	0.680
Open-Source LLMs													
Mistral7B	general	0.293	0.527	0.337	0.563	0.363	0.557	0.303	0.582	0.348	0.579	0.394	0.565
	structured	0.252	0.572	0.268	0.552	0.276	0.539	0.359	0.669	0.375	0.677	0.402	0.694
Mistral7B-FT	general	0.342	0.603	0.350	0.631	0.332	0.661	0.287	0.418	0.325	0.536	0.355	0.603
	structured	0.291	0.557	0.310	0.570	0.321	0.544	0.366	0.677	0.370	0.692	0.388	0.715
Mistral7B-MIMIC-FT	general	0.328	0.539	0.336	0.512	0.340	0.488	0.340	0.746	0.358	0.731	0.380	0.713
	structured	0.336	0.524	0.339	0.508	0.390	0.542	0.351	0.681	0.367	0.674	0.389	0.676
Llama3.1-8B	general	0.358	0.585	0.376	0.588	0.413	0.611	0.313	0.585	0.370	0.644	0.399	0.652
	structured	0.282	0.572	0.296	0.569	0.299	0.576	0.325	0.589	0.297	0.437	0.354	0.594
Llama3.1-8B-FT	general	0.343	0.605	0.373	0.593	0.404	0.621	0.312	0.589	0.360	0.620	0.403	0.624

	structured	0.289	0.558	0.301	0.571	0.294	0.575	0.332	0.589	0.353	0.630	0.386	0.649
Llama3.1-8B-MIMIC-FT	general	0.355	0.605	0.369	0.596	0.400	0.614	0.311	0.597	0.363	0.624	0.399	0.643
	structured	0.296	0.568	0.304	0.566	0.295	0.557	0.316	0.591	0.346	0.597	0.381	0.637
BioMistral7B	general	0.165	0.250	0.212	0.433	0.232	0.423	0.228	0.529	0.227	0.510	0.232	0.510
	structured	0.245	0.452	0.297	0.586	0.267	0.548	0.262	0.500	0.321	0.567	0.288	0.529
BioMistral7B-FT	general	0.254	0.439	0.272	0.455	0.165	0.231	0.273	0.460	0.286	0.462	0.311	0.468
	structured	0.287	0.436	0.361	0.542	0.306	0.594	0.353	0.563	0.360	0.563	0.264	0.483
Biomistral7B-MIMIC-FT	general	0.279	0.499	0.306	0.490	0.351	0.500	0.260	0.480	0.259	0.481	0.277	0.449
	structured	0.287	0.549	0.306	0.568	0.332	0.569	0.283	0.551	0.304	0.544	0.348	0.565

Table 2: We compared the performance of different models on zero-shot and few-shot tasks using top-3, top-5, and top-10 results for F1 and MRR across both closed- and open-source LLMs. The evaluation was conducted under three distinct settings: I. Vanilla models, including both closed- and open-source models. II. Models fine-tuned on a portion of the dataset. III. Models fine-tuned on 10,000 augmented data points generated from MIMIC-IV discharge notes. In many cases, open-source models continued to outperform closed-source models. The specific details of our experiments are provided in the Materials and Methods section. The complete scores, along with their 95% confidence intervals, are presented in Supplementary Table 5.

The highest-performing model that utilized our data augmentation strategy was Llama 3.1-8B-MIMIC-FT, achieving an F1 score of 0.400 (CI 95%: 0.414-0.387) in the top-10 medical jargon extractions using zero-shot prompts. The highest MRR score, 0.746 (CI 95%: 0.762-0.730), was achieved by Mistral7B-MIMIC-FT in the top-3 medical jargon extractions using few-shot prompts. However, neither the F1 nor the MRR score achieved the highest performance among all models that used the data augmentation strategy. In most cases, the F1 scores indicated that the data augmentation approach led to higher performance compared to vanilla models, and in some instances, even surpassed the performance of fine-tuned models. However, the performance differences varied depending on the prompt settings. These findings will be further discussed in the discussion section.

	top3		top5		top10	
	F1	MRR	F1	MRR	F1	MRR
GPT2[83]	0.097 (0.108/0.087)	0.191 (0.208/0.175)	0.094 (0.105/0.083)	0.193 (0.212/0.174)	0.099 (0.116/0.081)	0.170 (0.197/0.142)
BioGPT[92]	0.145 (0.163/0.126)	0.192 (0.235/0.149)	0.169 (0.197/0.141)	0.196 (0.212/0.180)	0.174 (0.202/0.146)	0.166 (0.219/0.112)

Table 3: Performance of Baseline models

Table 3 presents the performance of the baseline models, which include commonly used language models (LMs) for medical information extraction. As shown in Table 3, BioGPT [92] achieved the highest F1 score of 0.174 (95% CI: 0.202-0.146) and the highest MRR score of 0.196 (95% CI: 0.212-0.180) among the baseline models. However, these values are significantly lower than those achieved by the benchmark models, suggesting that LLMs generally outperform smaller models in downstream tasks.

Models	Prompt	Zero-Shot						Few-Shot					
		top3		top5		top10		top3		top5		top10	
		F1	MRR	F1	MRR	F1	MRR	F1	MRR	F1	MRR	F1	MRR
Mistral7B-10-MIMIC-FT	general	0.261	0.492	0.289	0.486	0.291	0.463	0.288	0.613	0.287	0.577	0.327	0.529
	structured	0.299	0.527	0.309	0.528	0.332	0.552	0.294	0.573	0.316	0.557	0.337	0.549
Mistral7B-100-MIMIC-FT	general	0.319	0.527	0.315	0.590	0.323	0.524	0.346	0.691	0.401	0.700	0.405	0.639
	structured	0.310	0.459	0.335	0.484	0.379	0.486	0.338	0.633	0.366	0.665	0.405	0.673
Mistral7B-1000-MIMIC-FT	general	0.331	0.531	0.348	0.516	0.336	0.465	0.336	0.746	0.358	0.702	0.367	0.726

	structured	0.311	0.484	0.341	0.481	0.398	0.535	0.348	0.693	0.377	0.672	0.381	0.672
Llama3.1-8B-10-MIMIC FT	general	0.335	0.602	0.381	0.583	0.395	0.610	0.304	0.610	0.352	0.614	0.394	0.634
	structured	0.288	0.559	0.300	0.578	0.304	0.580	0.331	0.618	0.358	0.626	0.379	0.632
Llama3.1-8B-100-MIMIC-FT	general	0.359	0.601	0.378	0.579	0.400	0.606	0.307	0.615	0.363	0.609	0.399	0.624
	structured	0.287	0.570	0.301	0.576	0.295	0.560	0.324	0.592	0.355	0.636	0.376	0.648
Llama3.1-8B-1000-MIMIC-FT	general	0.350	0.595	0.369	0.574	0.397	0.601	0.312	0.619	0.351	0.615	0.400	0.621
	structured	0.297	0.564	0.301	0.587	0.311	0.581	0.323	0.601	0.362	0.620	0.389	0.632
BioMistral7B-10-MIMIC-FT	general	0.270	0.459	0.295	0.453	0.321	0.464	0.241	0.473	0.257	0.488	0.276	0.464
	structured	0.300	0.539	0.302	0.535	0.301	0.517	0.213	0.403	0.237	0.445	0.241	0.427
BioMistral7B-100-MIMIC-FT	general	0.262	0.500	0.299	0.522	0.321	0.464	0.264	0.481	0.305	0.506	0.342	0.495
	structured	0.311	0.531	0.339	0.535	0.360	0.558	0.277	0.502	0.303	0.501	0.335	0.531
BioMistral7B-1000-MIMIC-FT	general	0.284	0.504	0.308	0.492	0.355	0.501	0.303	0.553	0.335	0.579	0.347	0.565
	structured	0.259	0.496	0.295	0.516	0.348	0.528	0.218	0.405	0.254	0.471	0.282	0.496

Table 4: Performance comparison between different open-source LLMs trained on different sizes of augmented data. We scale from 10 to 10,000 by multiplying 10. The results for using 10,000 are in Table 2. The F1 and MRR tends to increase when the size of the augmented dataset increases.

We also examined whether increasing or decreasing the size of the augmented dataset impacts the model's overall performance (Table 4). The highest F1 score of 0.405 was achieved by Mistral 7B when utilizing 100 augmented datasets with both general and specific prompts. The highest MRR score, 0.746, was observed from Mistral 7B fine-tuned on 10,000 augmented data points using a general prompt. A general trend was observed, indicating that utilizing larger augmented datasets results in higher performance compared to models trained on smaller datasets. As the dataset size increased, both the F1 and MRR scores improved; however, this trend was more pronounced for F1 than for MRR. In some cases, such as BioMistral7B in top-3 medical jargon extraction with zero-shot prompts, higher MRR scores were achieved with smaller dataset sizes.

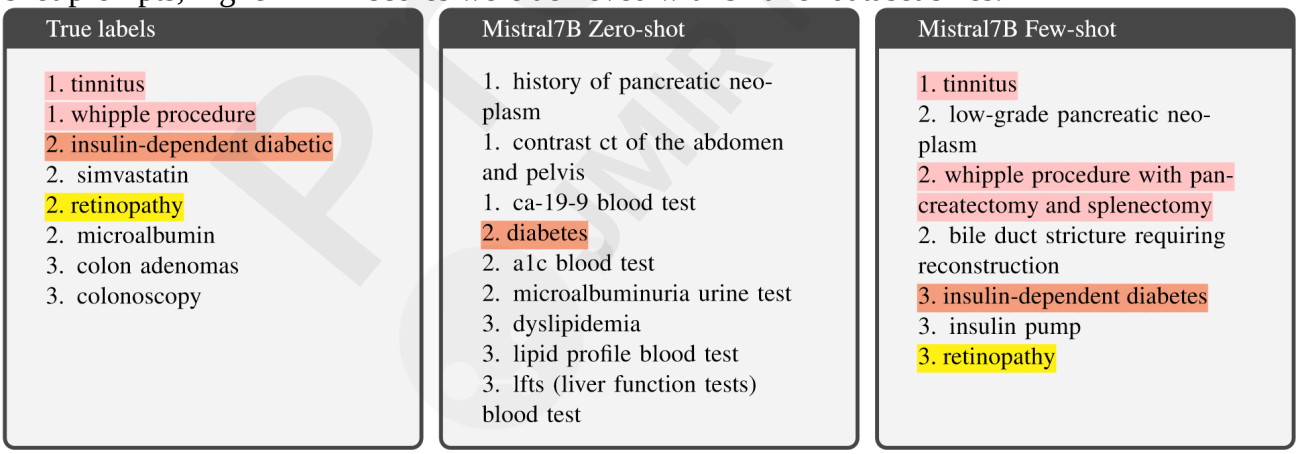

Figure 3: Case Study for Extracting the Top 3 Important Medical Jargons from Zero-shot and Few-shot Prompts in Mistral 7B. The few-shot prompting strategy demonstrates greater robustness in vanilla models compared to zero-shot prompting. The highlighted jargons represent terms that overlap with the expert-annotated labels, emphasizing the alignment between the model's outputs and domain experts' annotations.

Discussion

Zero-shot vs. Few-shot When it comes to prompting strategies, across both closed- and open-source models, fewshot prompts consistently outperformed zero-shot prompts in extracting relevant medical information (Table 2). As illustrated in Figure 3, the choice of prompting strategy can lead to different outputs, resulting in varying performance outcomes. This pattern aligns with findings from various studies, which have observed that few-shot prompts allow models to leverage examples that enhance contextual understanding and response specificity. For example, [89] demonstrated that few-shot prompting improves the performance of language models in complex tasks by providing contextual cues, enabling the model to better interpret nuanced queries—particularly in medical and technical domains where specificity is critical. However, in cases where the models were fine-tuned, these patterns began to diminish, with zero-shot prompts showing comparable or even higher scores than few-shot prompts. This can be attributed to instruction fine-tuning, which enhances a model's ability to handle unseen data. As noted by [95], instruction fine-tuning improves the performance of LLMs in zero-shot tasks by equipping them with a better understanding of task-specific instructions.

General vs. Structured Prompts The effectiveness of prompts can vary across models, with certain prompt styles enhancing performance for specific models (Table 2). While differences between models were generally minimal, some models performed better with particular prompt styles. Notably, for the Llama3.1 8B model, general prompts outperformed structured prompts in F1 and ranking prediction. Whereas in Mistral7B, structured prompts show improved performance to general prompts. This aligns with prior research suggesting that tailored prompts can optimize a model's performance in specialized tasks [96]. Thus, testing diverse prompt sets is essential to maximize model specific performance.

Finetuning LLMs Fine-tuning LLMs using domain-specific data proved effective in enhancing model performance, albeit with some limitations (Table 2). One possible explanation for this result is that the size of the datasets used for fine-tuning was too small to yield substantial performance gains. Increasing the size of the fine-tuning dataset is likely to further improve the model's performance [97]. Additionally, we observed that zero-shot prompts outperformed few-shot prompts, particularly in fine-tuned scenarios. Figure 4 illustrates the impact of fine-tuning: while vanilla models show limitations in extracting key points using prompts alone, fine-tuning enables the models to learn relevant patterns, leading to improved performance. Previous research has similarly shown that fine-tuning on domain-specific data can significantly enhance performance by adjusting model weights to reflect the unique characteristics of the target domain, such as better handling of abbreviations, acronyms, and clinically relevant contexts [59].

True Labels	BioMistral7B	BioMistral7B fintuned
1. tinnitus 1. whipple procedure 2. insulin-dependent diabetic 2. simvastatin 2. retinopathy 2. microalbumin 3. colon adenomas 3. colonoscopy 4. gerd 4. omeprazole 5. erectile dysfunction 5. viagra	1. tinnitus of many years' duration. 1. audiometry. 1. ct scan of the temporal bone. 2. low-grade pancreatic neoplasm. 2. ct scan of the abdomen and pelvis. 2. endoscopic ultrasound. 3. insulin-dependent diabetes. 3. fasting blood glucose. 3. hba1c. 3. c-peptide. ...	1. tinnitus of many years' duration. 2. low-grade pancreatic neoplasm. 3. insulin-dependent diabetes. 4. Dyslipidemia. 5. history of high blood pressure. 6. family history of brca gene. 7. multiple colon adenomas by colonoscopy. 8. sleep apnea. 9. Gerd. 10. erectile dysfunction.

Figure 4: Case Study for extracting Top 5 important medical jargons from BioMistral7B and BioMistral7B that was finetuned on some of the samples of the gold labeled dataset in Zero-shot prompt settings. The finetuned model shows more robustness than vanilla models, especially in Zero-shot prompts. The highlighted jargons are the ones that overlap with the expert-annotated labels.

Moreover, ICL has been shown to improve the model's zero-shot performance [95]. However, as the number of extracted terms increases, the performance gap between vanilla and fine-tuned models tends to narrow.

Data Augmentation We explored the impact of data augmentation using various sizes of MIMIC-IV [90] discharge notes to enhance the robustness of LLMs (Tables 2 and 4). Data augmentation simulated diverse scenarios within medical notes, enabling the LLM to generalize better across a broader range of note types. Similar to fine-tuning, the performance gains from data augmentation were marginal in few-shot prompt settings. However, in zero-shot prompts, the performance gain was significant, even exceeding the performance of models that used the fine-tuning strategy ($p < 0.05$). As shown in Figure 5, by utilizing only the augmented dataset, we achieved performance levels similar to those of fine-tuned models, and in some cases, even higher scores. These findings align with prior studies [95, 98], which argue that using LLM-generated data can improve model performance on downstream tasks. Additionally, studies have shown that while data augmentation provides benefits, substantial improvements often require a high degree of variation in the augmented data [99, 100]. In our case, data augmentation provided additional exposure to diverse linguistic patterns in medical notes. However, the approach alone was insufficient to dramatically outperform its counterparts, possibly due to the limited number of examples used in generating the augmented datasets.

Incrementing the size of augmented dataset As demonstrated in Table 4, increasing the size of the augmented dataset improves the overall performance of open-source LLMs, regardless of the model. This result aligns with findings from [101, 102], where increasing the size of synthetic datasets significantly enhanced model performance compared to models trained on smaller real-world datasets. However, we observed that MRR scores did not improve as much as F1 scores, indicating that ranking important information remains a non-trivial task for LLMs. As shown in Figure 5, while the model effectively captures important terms in the medical text, it still struggles to match the rankings as defined in the true labels. One possible explanation is the lower quality of the LLM-augmented dataset. Since we only used two examples for ICL and did not implement any filtering techniques to ensure the quality of the curated dataset, this may have affected the overall quality of the augmented dataset. We aim to address this limitation and improve performance in this aspect in future studies.

Our research demonstrated that mimicking expert-level human annotation is a non-trivial task, even for LLMs. While LLMs perform well in summarization and paraphrasing tasks, their recall drops significantly when it comes to identifying and prioritizing domain specific terms. Ranking identified information adds another layer of complexity. We

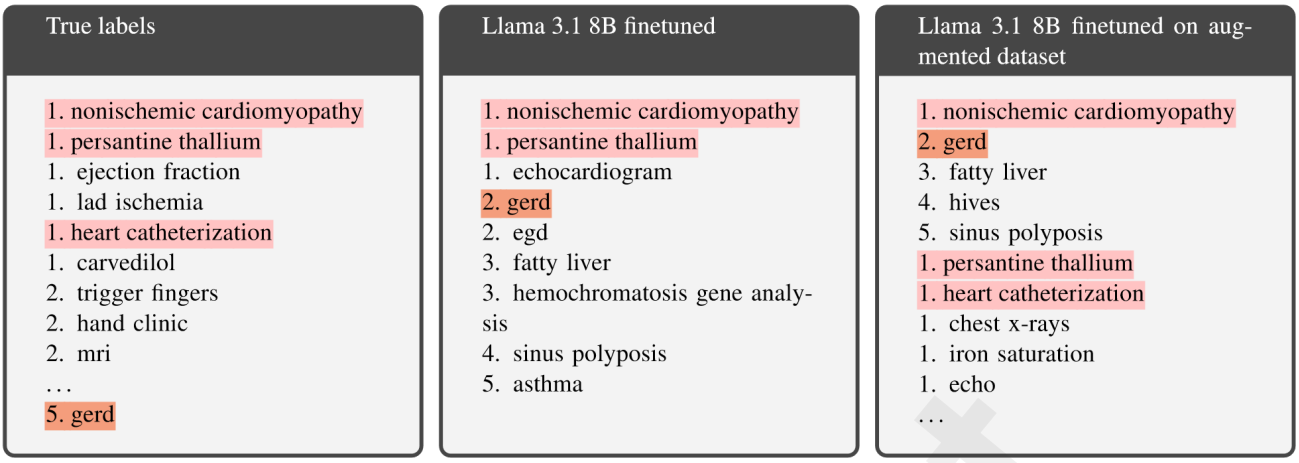

Figure 5: Case Study for extracting Top 5 important medical jargons from Llama3.1 8B finetuned and Llama 3.1 8B that was finetuned on the MIMIC-IV augmented dataset in Zero-shot prompt settings. In many cases, augmented models shows comparable performance than vanilla models and finetuned models, especially in Zero-shot prompts. The highlighted jargons are the ones that overlap with the expert annotated labels.

have shown that utilizing efficient fine-tuning and data augmentation can help improve model performance. Overall, our findings provide insights into the relative strengths and limitations of different methods for enhancing LLMs in medical applications.

Limitations

This study has several limitations. First, we tested only a limited selection of available closed- and open-source LLMs. Second, fine-tuning was not applied to closed-source LLMs. Third, despite the improvements achieved through finetuning and data augmentation, the performance of the models still falls short of human annotation. Future research will aim to address these challenges.

Conclusion

We evaluated both closed- and open-source LLMs for identifying and prioritizing medical jargon using expert-annotated EHR notes, demonstrating that fine-tuning and data augmentation significantly enhance performance. Comprehensive case analyses further validated our findings, highlighting the effectiveness of these methods for extracting and prioritizing important medical jargon for patients.

Acknowledgments

We greatly value UMass BioNLP group's insightful feedback and thoughtful guidance.

Author Contributions

Won Seok Jang, Sharmin Sultana, and Zonghai Yao authored the manuscript, with Won Seok and Sharmin implementing and analyzing models and Zonghai contributing to project planning.

Zhichao Yang, Hieu Tran, and Sunjae Kwon designed the framework and reviewed outcomes.

Hong Yu planned the project, wrote the proposal, and guided the research.

Conflict of Interest Statement

No conflicting interests.

DATA AVAILABILITY

The source code will be released here : https://www.github.com/memy85/2024 medicalnote annotation

References

1. Delbanco, T. *et al.* Open notes: doctors and patients signing on (2010).
2. HealthIT.gov. About the blue button movement. https://www.healthit.gov/patients-families/about-blue-button-movement (2024). [accessed 2024-10-29].
3. Delbanco, T. *et al.* Inviting patients to read their doctors' notes: a quasi-experimental study and a look ahead.
   Annals of internal medicine 157, 461–470 (2012).
4. Gabay, M. 21st century cures act. *Hospital pharmacy* 52, 264–265 (2017).
5. Bajwa, J., Munir, U., Nori, A. & Williams, B. Artificial intelligence in healthcare: transforming the practice of medicine. *Future healthcare journal* 8, e188 (2021).
6. Lye, C. T., Forman, H. P., Daniel, J. G. & Krumholz, H. M. The 21st century cures act and electronic health records one year later: will patients see the benefits? *Journal of the American Medical Informatics Association* 25, 1218–1220 (2018).
7. Arvisais-Anhalt, S. *et al.* The 21st century cures act and multiuser electronic health record access: potential pitfalls of information release. *Journal of medical Internet research* 24, e34085 (2022).
8. Rodriguez, J. A., Clark, C. R. & Bates, D. W. Digital health equity as a necessity in the 21st century cures act era. *Jama* 323, 2381–2382 (2020).
9. Nutbeam, D. Artificial intelligence and health literacy—proceed with caution. *Health Literacy and Communication Open* 1, 2263355 (2023).
10. Root, J. *et al.* Characteristics of patients who report confusion after reading their primary care clinic notes online. *Health communication* 31, 778–781 (2016).
11. Kayastha, N., Pollak, K. I. & LeBlanc, T. W. Open oncology notes: a qualitative study of oncology patients' experiences reading their cancer care notes. *Journal of Oncology Practice* 14, e251–e258 (2018).
12. Kujala, S. *et al.* Patients' experiences of web-based access to electronic health records in finland: Cross-sectional survey. *Journal of Medical Internet Research* 24, e37438 (2022).
13. Choudhry, A. J. *et al.* Readability of discharge summaries: with what level of information are we dismissing our patients? *The American Journal of Surgery* 211, 631–636 (2016).
14. Khasawneh, A., Kratzke, I., Adapa, K., Marks, L. & Mazur, L. Effect of notes' access and complexity on opennotes' utility. *Applied Clinical Informatics* 13, 1015–1023 (2022).
15. Rahimian, M. *et al.* Open notes sounds great, but will a provider's documentation change? an exploratory study of the effect of open notes on oncology documentation. *JAMIA open* 4, ooab051 (2021).
16. Zheng, J. & Yu, H. Readability formulas and user perceptions of electronic health records difficulty: a corpus study. *Journal of medical Internet research* 19, e59 (2017).
17. Zeng-Treitler, Q. *et al.* Text characteristics of clinical reports and their implications for the readability of personal health records. *Studies in health technology and informatics* 129, 1117 (2007).
18. Polepalli Ramesh, B., Houston, T., Brandt, C., Fang, H. & Yu, H. Improving patients' electronic health record comprehension with noteaid. In *MEDINFO 2013*, 714–718 (IOS

Press, 2013).
19. Sarzynski, E. *et al.* Opportunities to improve clinical summaries for patients at hospital discharge. *BMJ quality & safety* 26, 372–380 (2017).
20. Doak, C. C., Doak, L. G. & Root, J. H. Teaching patients with low literacy skills. *AJN The American Journal of Nursing* 96, 16M (1996).
21. Doak, C. C., Doak, L. G., Friedell, G. H. & Meade, C. D. Improving comprehension for cancer patients with low literacy skills: strategies for clinicians. *CA: A Cancer Journal for Clinicians* 48, 151–162 (1998).
22. Walsh, T. M. & Volsko, T. A. Readability assessment of internet-based consumer health information. *Respiratory care* 53, 1310–1315 (2008).
23. Eltorai, A. E., Han, A., Truntzer, J. & Daniels, A. H. Readability of patient education materials on the american orthopaedic society for sports medicine website. *The Physician and Sportsmedicine* 42, 125–130 (2014).
24. Morony, S., Flynn, M., McCaffery, K. J., Jansen, J. & Webster, A. C. Readability of written materials for ckd patients: a systematic review. *American Journal of Kidney Diseases* 65, 842–850 (2015).
25. Johnson, S. B., Farach, F. J., Pelphrey, K. & Rozenblit, L. Data management in clinical research: synthesizing stakeholder perspectives. *Journal of biomedical informatics* 60, 286–293 (2016).
26. Morid, M. A., Fiszman, M., Raja, K., Jonnalagadda, S. R. & Del Fiol, G. Classification of clinically useful sentences in clinical evidence resources. *Journal of biomedical informatics* 60, 14–22 (2016).
27. Kandula, S., Curtis, D. & Zeng-Treitler, Q. A semantic and syntactic text simplification tool for health content. In *AMIA annual symposium proceedings*, vol. 2010, 366 (American Medical Informatics Association, 2010).
28. Zeng-Treitler, Q., Goryachev, S., Kim, H., Keselman, A. & Rosendale, D. Making texts in electronic health records comprehensible to consumers: a prototype translator. In *AMIA Annual Symposium Proceedings*, vol. 2007, 846 (American Medical Informatics Association, 2007).
29. Abrahamsson, E., Forni, T., Skeppstedt, M. & Kvist, M. Medical text simplification using synonym replacement: Adapting assessment of word difficulty to a compounding language. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, 57–65 (2014).
30. Zheng, J. & Yu, H. Methods for linking ehr notes to education materials. *Information Retrieval Journal* 19, 174–188 (2016).
31. Chen, J. *et al.* A natural language processing system that links medical terms in electronic health record notes to lay definitions: system development using physician reviews. *Journal of medical Internet research* 20, e26 (2018).
32. Kwon, S. *et al.* Medjex: A medical jargon extraction model with wiki's hyperlink span and contextualized masked language model score. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, vol. 2022, 11733 (NIH Public Access, 2022).
33. Leroy, G., Endicott, J. E., Mouradi, O., Kauchak, D. & Just, M. L. Improving perceived and actual text difficulty for health information consumers using semi-automated methods. In *AMIA Annual Symposium Proceedings*, vol. 2012, 522 (American Medical Informatics Association, 2012).
34. Chen, J., Zheng, J. & Yu, H. Finding Important Terms for Patients in Their Electronic Health Records: A Learning-to-Rank Approach Using Expert Annotations 4, e6373. URL https://medinform.jmir.org/2016/4/e40.
35. Chen, J. & Yu, H. Unsupervised ensemble ranking of terms in electronic health record notes

based on their importance to patients 68, 121–131. URL https://www.sciencedirect.com/science/article/pii/S153204641730045X.

36. Aronson, A. R. Metamap: Mapping text to the umls metathesaurus. *Bethesda, MD: NLM, NIH, DHHS* 1, 26 (2006).
37. Neumann, M., King, D., Beltagy, I. & Ammar, W. Scispacy: fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669* (2019).
38. Eyre, H. *et al.* Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. *AMIA Annu Symp Proc* 2021, 438–447 (2021).
39. Soldaini, L. & Goharian, N. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir*, 1–4 (2016).
40. Unified Medical Language System® (UMLS®) – Basics .
41. Tian, S. *et al.* Opportunities and challenges for chatgpt and large language models in biomedicine and health. *Briefings in Bioinformatics* 25, bbad493 (2024).
42. Singhal, K. *et al.* Large language models encode clinical knowledge. *Nature* 620, 172–180 (2023).
43. Singhal, K. *et al.* Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617* (2023).
44. Tu, T. *et al.* Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654* (2024).
45. McDuff, D. *et al.* Towards accurate differential diagnosis with large language models. *arXiv preprint arXiv:2312.00164* (2023).
46. Wu, C. *et al.* Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association* ocae045 (2024).
47. Chen, Z. *et al.* Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079* (2023).
48. Tran, H., Yang, Z., Yao, Z. & Yu, H. Bioinstruct: Instruction tuning of large language models for biomedical natural language processing. *arXiv preprint arXiv:2310.19975* (2023).
49. Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375* (2023).
50. Kung, T. H. *et al.* Performance of chatgpt on usmle: potential for ai-assisted medical education using large language models. *PLoS digital health* 2, e0000198 (2023).
51. Yang, L. *et al.* Advancing multimodal medical capabilities of gemini. *arXiv preprint arXiv:2405.03162* (2024).
52. Yang, Z. *et al.* Performance of multimodal gpt-4v on usmle with image: potential for imaging diagnostic support with explanations. *medRxiv* 2023–10 (2023).
53. Yao, Z. *et al.* Medqa-cs: Benchmarking large language models clinical skills using an ai-sce framework. *arXiv preprint arXiv:2410.01553* (2024).
54. Hu, Y. *et al.* Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association* ocad259 (2024).
55. Monajatipoor, M. *et al.* LLMs in Biomedicine: A study on clinical Named Entity Recognition. URL http: //arxiv.org/abs/2404.07376. 2404.07376.
56. Hu, D., Liu, B., Zhu, X., Lu, X. & Wu, N. Zero-shot information extraction from radiological reports using chatgpt. *International Journal of Medical Informatics* 183, 105321 (2024).
57. Liu, S., Wang, A., Xiu, X., Zhong, M. & Wu, S. Evaluating Medical Entity Recognition in Health Care: Entity Model Quantitative Study 12, e59782. URL https://medinform.jmir.org/2024/1/e59782.
58. Bose, P. *et al.* A Survey on Recent Named Entity Recognition and Relationship Extraction Techniques on Clinical Texts 11, 8319. URL https://www.mdpi.com/2076-3417/11/18/8319.
59. Lee, J. *et al.* Biobert: a pre-trained biomedical language representation model for biomedical

text mining. *Bioinformatics* 36, 1234–1240 (2020).
60. Liu, Y. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
61. Yao, Z., Cao, Y., Yang, Z., Deshpande, V. & Yu, H. Extracting biomedical factual knowledge using pretrained language model and electronic health record context. In *AMIA Annual Symposium Proceedings*, vol. 2022, 1188 (2023).
62. Yao, Z., Cao, Y., Yang, Z. & Yu, H. Context variance evaluation of pretrained language models for prompt-based biomedical knowledge probing. *AMIA Summits on Translational Science Proceedings* 2023, 592 (2023).
63. Gutierrez, B. J. *et al.* Thinking about gpt-3 in-context learning for biomedical ie? think again. *arXiv preprint arXiv:2203.08410* (2022).
64. Moradi, M., Blagec, K., Haberl, F. & Samwald, M. Gpt-3 models are poor few-shot learners in the biomedical domain. *arXiv preprint arXiv:2109.02555* (2021).
65. Devlin, J. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
66. Alsentzer, E. *et al.* Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323* (2019).
67. Lim, J. H., Kwon, S., Yao, Z., Lalor, J. P. & Yu, H. Large language model-based role-playing for personalized medical jargon extraction. *arXiv preprint arXiv:2408.05555* (2024).
68. OpenAI. Openai website. URL https://openai.com/.
69. Ghali, M.-K. *et al.* Gamedx: Generative ai-based medical entity data extractor using large language models. *arXiv preprint arXiv:2405.20585* (2024).
70. Butler, J. J. *et al.* From jargon to clarity: Improving the readability of foot and ankle radiology reports with an artificial intelligence large language model. *Foot and Ankle Surgery* 30, 331–337 (2024).
71. Mannhardt, N. *et al.* Impact of large language model assistance on patients reading clinical notes: A mixedmethods study. *arXiv preprint arXiv:2401.09637* (2024).
72. Lu, J., Li, J., Wallace, B. C., He, Y. & Pergola, G. Napss: Paragraph-level medical text simplification via narrative prompting and sentence-matching summarization. *arXiv preprint arXiv:2302.05574* (2023).
73. Speier, W., Ong, M. K. & Arnold, C. W. Using phrases and document metadata to improve topic modeling of clinical reports 61, 260–266. URL https://www.sciencedirect.com/science/article/pii/S1532046416300284.
74. Wen, Z. *et al.* Mining heterogeneous clinical notes by multi-modal latent topic model 16, e0249622 (2021. 4.
8.). URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0249622.
75. Sun, S., Zack, T., Williams, C. Y. K., Sushil, M. & Butte, A. J. Topic modeling on clinical social work notes for exploring social determinants of health factors 7, ooad112. URL https://www.ncbi.nlm.nih.gov/pmc/articles/ PMC10788143/. 38223407.
76. Chen, J., Jagannatha, A. N., Fodeh, S. J. & Yu, H. Ranking Medical Terms to Support Expansion of Lay Language Resources for Patient Comprehension of Electronic Health Record Notes: Adapted Distant Supervision Approach 5, e42. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5686421/. 29089288.
77. Aronson, A. R. & Lang, F.-M. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association* 17, 229–236 (2010).
78. Yao, Z. *et al.* Readme: Bridging medical jargon and lay understanding for patient education through data-centric nlp. *arXiv preprint arXiv:2312.15561* (2023).
79. Cai, P. *et al.* Generation of patient after-visit summaries to support physicians. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING)* (2022).

80. Jiang, A. Q. *et al.* Mistral 7B. URL http://arxiv.org/abs/2310.06825. 2310.06825.
81. Labrak, Y. *et al.* BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. URL http://arxiv.org/abs/2402.10373. 2402.10373.
82. Dubey, A. *et al.* The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).
83. OpenAI, R. *et al.* Gpt-4 technical report. *ArXiv* 2303, 08774 (2023).
84. Ye, J. *et al.* A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv preprint arXiv:2303.10420* (2023).
85. OpenAI. Gpt-4o mini: Advancing cost-efficient intelligence (2024). URL https://openai.com/index/ gpt-4o-mini-advancing-cost-efficient-intelligence/.
86. Anthropic. Introduction to claude. URL https://docs.anthropic.com/en/docs/intro-to-claude.
87. Hu, E. J. *et al.* LoRA: Low-Rank Adaptation of Large Language Models. URL http://arxiv.org/abs/2106.09685. 2106.09685.
88. Li, R., Wang, X. & Yu, H. Two Directions for Clinical Data Generation with Large Language Models: Data-toLabel and Label-to-Data 2023, 7129–7143. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10782150/. 38213944.
89. Brown, T. B. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).
90. Johnson, A. E. W. *et al.* MIMIC-IV, a freely accessible electronic health record dataset 10, 1. URL https: //www.nature.com/articles/s41597-022-01899-x.
91. Radford, A. *et al.* Language models are unsupervised multitask learners. *OpenAI blog* 1, 9 (2019).
92. Luo, R. *et al.* Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics* 23, bbac409 (2022).
93. Turney, P. D. Learning algorithms for keyphrase extraction. *Information retrieval* 2, 303–336 (2000).
94. Zesch, T. & Gurevych, I. Approximate Matching for Evaluating Keyphrase Extraction. In Angelova, G. & Mitkov, R. (eds.) *Proceedings of the International Conference RANLP-2009*, 484–489 (Association for Computational Linguistics). URL https://aclanthology.org/R09-1086.
95. Wei, J. *et al.* Finetuned Language Models Are Zero-Shot Learners. URL http://arxiv.org/abs/2109.01652. 2109. 01652.
96. Liu, J. *et al.* What Makes Good In-Context Examples for GPT-3? URL http://arxiv.org/abs/ 2101.06804. 2101.06804.
97. Vieira, I., Allred, W., Lankford, S., Castilho, S. & Way, A. How Much Data is Enough Data? Fine-Tuning Large Language Models for In-House Translation: Performance Evaluation Across Multiple Dataset Sizes. URL http://arxiv.org/abs/2409.03454. 2409.03454.
98. Tang, R., Han, X., Jiang, X. & Hu, X. Does Synthetic Data Generation of LLMs Help Clinical Text Mining? URL http://arxiv.org/abs/2303.04360. 2303.04360.
99. Wei, J. & Zou, K. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. URL http://arxiv.org/abs/1901.11196. 1901.11196.
100. Chen, J., Tam, D., Raffel, C., Bansal, M. & Yang, D. An Empirical Survey of Data Augmentation for Limited Data Learning in NLP. URL http://arxiv.org/abs/2106.07499. 2106.07499.
101. Yuan, J., Zhang, J., Sun, S., Torr, P. & Zhao, B. Real-Fake: Effective Training Data Synthesis Through Distribution Matching. URL http://arxiv.org/abs/2310.10402. 2310.10402.
102. Kim, H. *et al.* Synthetic Data Improve Survival Status Prediction Models in Early-Onset Colorectal Cancer 8, e2300201. 38271642.

Appendix

Performance Analysis on F1, MRR, Precision and Recall Score

Close-Source and Open-Source Models

Models	Prompt	Zero-Shot						Few-Shot					
		top3		top5		top10		top3		top5		top10	
		F1	MRR	F1	MRR	F1	MRR	F1	MRR	F1	MRR	F1	MRR
Close-Source LLMs													
GPT-4 Turbo	general	0.306±0.009	0.562±0.033	0.359±0.016	0.601±0.029	0.432±0.007	0.640±0.04	0.324±0.018	0.600±0.0325	0.368±0.013	0.626±0.032	0.433±0.015	0.621±0.0305
	specific	0.317±0.012	0.643±0.0265	0.347±0.016	0.668±0.04	0.370±0.0085	0.691±0.0305	0.362±0.0155	0.673±0.0245	0.398±0.0135	0.705±0.0125	0.424±0.0135	0.691±0.0255
GPT-4o Mini	general	0.320±0.015	0.617±0.026	0.355±0.016	0.667±0.026	0.392±0.008	0.691±0.014	0.329±0.007	0.664±0.015	0.369±0.0075	0.700±0.0335	0.383±0.012	0.708±0.049
	specific	0.303±0.012	0.620±0.0185	0.329±0.0115	0.617±0.0245	0.335±0.006	0.610±0.041	0.328±0.017	0.666±0.0285	0.363±0.0085	0.668±0.0315	0.381±0.0075	0.684±0.023
GPT-3.5 Turbo	general	0.261±0.018	0.478±0.023	0.308±0.006	0.582±0.0155	0.324±0.0055	0.602±0.026	0.343±0.0155	0.649±0.029	0.359±0.009	0.655±0.0245	0.378±0.0125	0.669±0.02
	specific	0.303±0.0035	0.575±0.0045	0.329±0.0045	0.573±0.0115	0.343±0.013	0.572±0.021	0.354±0.0195	0.661±0.017	0.367±0.017	0.633±0.0385	0.398±0.01	0.661±0.0055
Claude 3.0 Sonnet	general	0.289±0.003	0.554±0.016	0.328±0.0075	0.556±0.029	0.355±0.0095	0.557±0.033	0.308±0.0095	0.596±0.028	0.345±0.0075	0.598±0.0305	0.399±0.0065	0.639±0.027
	specific	0.304±0.009	0.666±0.024	0.317±0.007	0.673±0.013	0.323±0.0115	0.668±0.027	0.354±0.013	0.683±0.026	0.357±0.012	0.684±0.0145	0.371±0.0105	0.692±0.021
Claude 3.0 Haiku	general	0.284±0.0065	0.613±0.0255	0.320±0.0065	0.629±0.015	0.345±0.013	0.620±0.02	0.320±0.025	0.672±0.023	0.358±0.009	0.690±0.018	0.387±0.0075	0.683±0.0235
	specific	0.248±0.011	0.584±0.0315	0.271±0.009	0.591±0.0265	0.281±0.0065	0.614±0.0275	0.357±0.0115	0.668±0.018	0.364±0.0135	0.715±0.016	0.349±0.0065	0.693±0.0185
Claude 3.0 Opus	general	0.340±0.015	0.707±0.034	0.365±0.0135	0.698±0.0185	0.390±0.0055	0.713±0.023	0.316±0.019	0.615±0.027	0.359±0.0085	0.661±0.019	0.412±0.011	0.709±0.023
	specific	0.289±0.009	0.617±0.027	0.344±0.011	0.643±0.0295	0.387±0.008	0.664±0.035	0.319±0.0115	0.646±0.0275	0.331±0.0135	0.551±0.013	0.387±0.006	0.680±0.013
Open-Source LLMs													
Mistral-7b	general	0.293±0.0115	0.527±0.029	0.337±0.0155	0.563±0.0275	0.363±0.0075	0.557±0.0255	0.303±0.0155	0.582±0.02	0.348±0.0075	0.579±0.026	0.394±0.01	0.565±0.0285
	specific	0.252±0.0095	0.572±0.0345	0.268±0.0085	0.552±0.029	0.276±0.008	0.539±0.022	0.359±0.011	0.669±0.0195	0.375±0.0105	0.677±0.017	0.402±0.0065	0.694±0.019
Mistral-7b-FT	general	0.342±0.0065	0.603±0.0165	0.350±0.0075	0.631±0.0035	0.332±0.008	0.661±0.019	0.287±0.008	0.418±0.018	0.325±0.007	0.536±0.0255	0.355±0.004	0.603±0.015
	specific	0.291±0.023	0.557±0.033	0.310±0.017	0.570±0.0295	0.321±0.0085	0.544±0.026	0.366±0.017	0.677±0.0185	0.370±0.0105	0.692±0.0115	0.388±0.023	0.715±0.019
Mistral-7b-MIMIC-FT	general	0.328±0.003	0.539±0.023	0.336±0.010	0.512±0.026	0.340±0.009	0.488±0.030	0.340±0.007	0.746±0.016	0.358±0.008	0.731±0.008	0.380±0.008	0.713±0.016
	specific	0.336±0.012	0.524±0.032	0.339±0.008	0.508±0.030	0.390±0.005	0.542±0.023	0.351±0.008	0.681±0.014	0.367±0.007	0.674±0.022	0.389±0.006	0.676±0.029
Llama 3.1-8b	general	0.358±0.018	0.585±0.017	0.376±0.015	0.588±0.017	0.413±0.013	0.611±0.020	0.313±0.010	0.585±0.014	0.370±0.008	0.644±0.037	0.399±0.011	0.652±0.036
	specific	0.282±0.015	0.572±0.033	0.296±0.013	0.569±0.021	0.299±0.005	0.576±0.029	0.325±0.006	0.589±0.030	0.297±0.010	0.437±0.016	0.354±0.010	0.594±0.029
Llama 3.1-8b-FT	general	0.343±0.015	0.605±0.022	0.373±0.007	0.593±0.016	0.404±0.008	0.621±0.020	0.312±0.013	0.589±0.024	0.360±0.009	0.620±0.021	0.403±0.012	0.624±0.013
	specific	0.289±0.008	0.558±0.016	0.301±0.010	0.571±0.024	0.294±0.012	0.575±0.022	0.332±0.012	0.589±0.033	0.353±0.009	0.630±0.031	0.386±0.012	0.649±0.024
llama 3.1-8b-MIMIC-FT	general	0.355±0.020	0.605±0.023	0.369±0.008	0.596±0.021	0.400±0.013	0.614±0.016	0.311±0.010	0.597±0.021	0.363±0.009	0.624±0.025	0.399±0.012	0.643±0.026
	specific	0.296±0.015	0.568±0.031	0.304±0.010	0.566±0.036	0.295±0.005	0.557±0.042	0.316±0.007	0.591±0.020	0.346±0.013	0.597±0.034	0.381±0.009	0.637±0.041
Biomistral-7b	general	0.165±0.013	0.250±0.030	0.212±0.021	0.433±0.036	0.232±0.013	0.423±0.024	0.228±0.005	0.529±0.029	0.227±0.005	0.510±0.028	0.232±0.007	0.510±0.024
	specific	0.245±0.012	0.452±0.009	0.297±0.013	0.586±0.022	0.267±0.013	0.548±0.017	0.262±0.007	0.500±0.037	0.321±0.014	0.567±0.026	0.288±0.018	0.529±0.034
Biomistral-7b-FT	general	0.254±0.011	0.439±0.036	0.272±0.019	0.455±0.021	0.165±0.039	0.231±0.052	0.273±0.017	0.460±0.039	0.286±0.014	0.462±0.034	0.311±0.011	0.468±0.020
	specific	0.287±0.061	0.436±0.101	0.361±0.034	0.542±0.048	0.306±0.026	0.594±0.061	0.353±0.011	0.563±0.035	0.360±0.021	0.563±0.023	0.264±0.021	0.483±0.038
Biomistral-7b-MIMIC-FT	general	0.279±0.010	0.499±0.030	0.306±0.007	0.490±0.038	0.351±0.012	0.500±0.036	0.260±0.004	0.480±0.016	0.259±0.007	0.481±0.025	0.277±0.006	0.449±0.020
	specific	0.287±0.010	0.549±0.026	0.306±0.015	0.568±0.025	0.332±0.007	0.569±0.033	0.283±0.007	0.551±0.026	0.304±0.005	0.544±0.023	0.348±0.007	0.565±0.019

Table 5: Performance comparison between different models on zero-shot and few-shot tasks with top-3, top-5, and top-10 results with 95% confidence interval for F1 and MRR in closed and open source LLMs.

Models	Prompt	Zero-Shot						Few-Shot					
		top3		top5		top10		top3		top5		top10	
		P	R	P	R	P	R	P	R	P	R	P	R
Close-Source LLMs													
GPT-4 Turbo	general	0.449±0.012	0.231±0.007	0.445±0.013	0.300±0.018	0.404±0.016	0.463±0.014	0.489±0.023	0.243±0.015	0.478±0.010	0.300±0.014	0.415±0.013	0.452±0.022
	specific	0.296±0.011	0.343±0.016	0.305±0.014	0.403±0.023	0.282±0.007	0.539±0.017	0.345±0.013	0.381±0.020	0.354±0.010	0.455±0.021	0.335±0.010	0.579±0.021

GPT-4o Mini	general	0.449±0.021 0.249±0.013	0.430±0.012 0.303±0.017	0.356±0.004 0.437±0.018	0.358±0.011 0.304±0.007	0.389±0.007 0.351±0.010	0.338±0.006 0.441±0.013
	specific	0.283±0.008 0.325±0.017	0.289±0.011 0.383±0.018	0.257±0.003 0.482±0.021	0.300±0.017 0.362±0.019	0.311±0.007 0.437±0.018	0.306±0.009 0.508±0.020
GPT-3.5 Turbo	general	0.406±0.023 0.192±0.014	0.370±0.010 0.265±0.011	0.281±0.008 0.384±0.003	0.361±0.018 0.327±0.019	0.358±0.008 0.361±0.012	0.350±0.008 0.412±0.023
	specific	0.284±0.008 0.324±0.007	0.285±0.004 0.390±0.010	0.267±0.011 0.478±0.020	0.349±0.015 0.360±0.024	0.335±0.016 0.406±0.022	0.342±0.008 0.476±0.016
Claude 3.0 Sonnet	general	0.391±0.010 0.229±0.002	0.375±0.010 0.292±0.007	0.326±0.007 0.389±0.016	0.440±0.011 0.237±0.011	0.421±0.005 0.293±0.011	0.371±0.010 0.432±0.007
	specific	0.272±0.012 0.346±0.013	0.265±0.003 0.395±0.017	0.238±0.011 0.502±0.011	0.330±0.018 0.382±0.010	0.307±0.011 0.426±0.015	0.285±0.013 0.530±0.015
Claude 3.0 Haiku	general	0.398±0.014 0.220±0.004	0.395±0.017 0.270±0.005	0.313±0.013 0.384±0.013	0.463±0.023 0.245±0.024	0.458±0.009 0.294±0.011	0.365±0.004 0.413±0.015
	specific	0.222±0.010 0.281±0.013	0.225±0.010 0.341±0.010	0.210±0.006 0.425±0.008	0.322±0.016 0.399±0.010	0.302±0.010 0.459±0.022	0.257±0.007 0.544±0.021
Claude 3.0 Opus	general	0.386±0.013 0.271±0.012	0.373±0.011 0.297±0.016	0.337±0.006 0.455±0.016	0.438±0.025 0.249±0.028	0.375±0.011 0.344±0.008	0.359±0.012 0.486±0.019
	specific	0.273±0.007 0.307±0.015	0.309±0.012 0.389±0.016	0.309±0.010 0.518±0.009	0.312±0.011 0.374±0.025	0.320±0.013 0.426±0.016	0.306±0.007 0.538±0.013
Open-Source LLMs							
Mistral-7b	general	0.431±0.014 0.223±0.011	0.376±0.018 0.305±0.015	0.333±0.005 0.398±0.013	0.455±0.019 0.227±0.013	0.433±0.010 0.291±0.008	0.385±0.013 0.404±0.008
	specific	0.216±0.007 0.303±0.015	0.215±0.007 0.355±0.013	0.197±0.007 0.460±0.008	0.351±0.009 0.367±0.015	0.322±0.010 0.449±0.013	0.301±0.005 0.604±0.008
Mistral-7b-FT	general	0.394±0.014 0.303±0.004	0.329±0.005 0.374±0.018	0.276±0.005 0.416±0.015	0.311±0.014 0.267±0.006	0.312±0.007 0.340±0.015	0.335±0.007 0.378±0.008
	specific	0.251±0.019 0.345±0.030	0.253±0.013 0.401±0.028	0.236±0.006 0.505±0.017	0.357±0.018 0.375±0.017	0.332±0.010 0.419±0.014	0.310±0.012 0.520±0.009
Mistral-7b-MIMIC-FT	general	0.308±0.011 0.351±0.013	0.269±0.012 0.448±0.013	0.305±0.013 0.387±0.018	0.261±0.007 0.488±0.014	0.281±0.010 0.497±0.006	0.304±0.009 0.507±0.011
	specific	0.376±0.017 0.304±0.013	0.351±0.006 0.327±0.011	0.345±0.006 0.449±0.011	0.325±0.006 0.383±0.014	0.310±0.004 0.450±0.015	0.308±0.007 0.528±0.014
Llama 3.1-8b	general	0.301±0.018 0.444±0.035	0.334±0.012 0.431±0.023	0.330±0.016 0.551±0.017	0.286±0.011 0.346±0.010	0.337±0.010 0.409±0.010	0.331±0.010 0.503±0.018
	specific	0.231±0.010 0.362±0.023	0.234±0.009 0.403±0.022	0.205±0.002 0.551±0.014	0.308±0.004 0.345±0.013	0.297±0.010 0.437±0.016	0.287±0.008 0.585±0.026
Llama 3.1-8b-FT	general	0.287±0.019 0.428±0.019	0.326±0.011 0.436±0.017	0.320±0.012 0.550±0.028	0.286±0.015 0.344±0.014	0.332±0.003 0.393±0.019	0.335±0.007 0.507±0.025
	specific	0.238±0.006 0.367±0.017	0.235±0.007 0.417±0.017	0.202±0.009 0.536±0.020	0.312±0.010 0.355±0.017	0.295±0.007 0.439±0.015	0.286±0.010 0.592±0.023
llama 3.1-8b-MIMIC-FT	general	0.300±0.025 0.439±0.023	0.328±0.010 0.423±0.016	0.322±0.015 0.529±0.015	0.283±0.011 0.346±0.009	0.329±0.007 0.405±0.016	0.332±0.015 0.498±0.008
	specific	0.246±0.013 0.371±0.016	0.240±0.007 0.415±0.020	0.203±0.005 0.540±0.005	0.298±0.012 0.338±0.014	0.288±0.011 0.434±0.019	0.284±0.008 0.579±0.009
Biomistral-7b	general	0.336±0.035 0.109±0.008	0.536±0.040 0.132±0.013	0.548±0.018 0.147±0.010	0.654±0.025 0.137±0.005	0.683±0.019 0.136±0.003	0.673±0.015 0.140±0.005
	specific	0.538±0.013 0.158±0.010	0.692±0.017 0.189±0.009	0.673±0.022 0.166±0.009	0.673±0.024 0.163±0.006	0.740±0.024 0.205±0.010	0.721±0.028 0.180±0.012
Biomistral-7b-FT	general	0.341±0.033 0.204±0.018	0.330±0.034 0.231±0.012	0.167±0.042 0.164±0.038	0.278±0.020 0.272±0.039	0.321±0.019 0.259±0.018	0.317±0.020 0.306±0.011
	specific	0.471±0.116 0.207±0.041	0.603±0.060 0.258±0.025	0.288±0.023 0.329±0.040	0.592±0.033 0.252±0.011	0.580±0.055 0.262±0.012	0.278±0.023 0.256±0.041
Biomistral-7b-MIMIC-FT	general	0.428±0.015 0.207±0.008	0.398±0.005 0.249±0.008	0.361±0.013 0.343±0.012	0.276±0.005 0.246±0.007	0.265±0.006 0.253±0.008	0.285±0.009 0.270±0.005
	specific	0.329±0.010 0.255±0.010	0.344±0.014 0.277±0.015	0.324±0.011 0.340±0.007	0.367±0.008 0.230±0.009	0.360±0.007 0.264±0.006	0.338±0.007 0.358±0.009

Table 6: Performance comparison between different models on zero-shot and few-shot tasks with top-3, top-5, and top-10 results with 95% confidence interval for Precision (P) and Recall (R) in closed and open source LLMs.

Baseline Models

	top3		top5		top10	
	P	R	P	R	P	R
GPT2[83]	0.102 (0.111/0.093)	0.094 (0.109/0.080)	0.108 (0.119/0.097)	0.085 (0.099/0.070)	0.095 (0.113/0.076)	0.105 (0.124/0.085)
BioGPT[92]	0.339 (0.393/0.284)	0.092 (0.103/0.081)	0.411 (0.464/0.358)	0.106 (0.125/0.088)	0.414 (0.464/0.363)	0.110 (0.130/0.090)

Table 7: Precision (P) and Recall (R) for Baseline models

Impact of Augmented Data Scale on Open-Source Language Model Training

		Zero-Shot			Few-Shot		
Models	Prompt	top3	top5	top10	top3	top5	top10

		P	R	P	R	P	R	P	R	P	R	P	R
Open-Source LLMs													
Mistral 7b 10MIMIC FT	general	0.367±0.020	0.202±0.011	0.337±0.014	0.253±0.013	0.278±0.004	0.305±0.016	0.362±0.019	0.240±0.010	0.344±0.014	0.246±0.011	0.349±0.011	0.308±0.011
	specific	0.336±0.015	0.270±0.012	0.344±0.012	0.281±0.011	0.321±0.013	0.343±0.011	0.396±0.014	0.233±0.008	0.376±0.012	0.273±0.009	0.365±0.009	0.314±0.010
Mistral 7b 100MIMIC FT	general	0.307±0.020	0.332±0.010	0.243±0.014	0.450±0.008	0.289±0.013	0.367±0.009	0.314±0.017	0.386±0.008	0.380±0.010	0.423±0.014	0.369±0.018	0.451±0.015
	specific	0.371±0.017	0.267±0.014	0.350±0.012	0.321±0.017	0.361±0.009	0.399±0.006	0.339±0.008	0.337±0.011	0.321±0.006	0.425±0.013	0.338±0.005	0.506±0.016
Mistral 7b 1000MIMIC FT	general	0.314±0.013	0.350±0.008	0.280±0.006	0.460±0.012	0.298±0.008	0.384±0.018	0.253±0.010	0.501±0.011	0.287±0.011	0.475±0.004	0.293±0.010	0.492±0.012
	specific	0.376±0.017	0.304±0.013	0.351±0.006	0.327±0.011	0.345±0.006	0.449±0.011	0.325±0.006	0.383±0.014	0.310±0.004	0.450±0.015	0.308±0.007	0.528±0.014
Llama 3.1 8b 10MIMIC FT	general	0.271±0.014	0.439±0.033	0.339±0.013	0.435±0.015	0.320±0.007	0.516±0.016	0.280±0.010	0.333±0.011	0.322±0.003	0.389±0.015	0.328±0.008	0.496±0.010
	specific	0.239±0.014	0.362±0.016	0.235±0.012	0.413±0.018	0.211±0.004	0.547±0.022	0.308±0.014	0.360±0.028	0.300±0.014	0.445±0.013	0.283±0.009	0.578±0.014
Llama 3.1 8b 100MIMIC FT	general	0.303±0.017	0.439±0.007	0.337±0.014	0.433±0.013	0.321±0.013	0.530±0.016	0.283±0.020	0.335±0.016	0.332±0.011	0.400±0.006	0.330±0.004	0.507±0.021
	specific	0.240±0.011	0.357±0.017	0.237±0.003	0.413±0.014	0.204±0.004	0.529±0.013	0.302±0.008	0.351±0.025	0.297±0.009	0.443±0.013	0.278±0.005	0.580±0.016
Llama 3.1 8b 1000MIMIC FT	general	0.285±0.013	0.454±0.017	0.332±0.016	0.416±0.021	0.324±0.009	0.513±0.021	0.287±0.012	0.340±0.020	0.324±0.014	0.384±0.009	0.335±0.009	0.497±0.009
	specific	0.244±0.012	0.381±0.017	0.238±0.009	0.409±0.006	0.216±0.005	0.553±0.012	0.303±0.006	0.346±0.005	0.305±0.012	0.444±0.024	0.288±0.010	0.598±0.016
BioMistral 7b 10MIMIC FT	general	0.398±0.015	0.204±0.007	0.369±0.008	0.246±0.007	0.331±0.009	0.313±0.005	0.249±0.006	0.233±0.011	0.266±0.008	0.248±0.008	0.283±0.006	0.269±0.009
	specific	0.358±0.010	0.258±0.010	0.333±0.006	0.277±0.011	0.298±0.004	0.304±0.007	0.285±0.020	0.170±0.011	0.291±0.011	0.200±0.011	0.296±0.012	0.204±0.006
BioMistral 7b 100MIMIC FT	general	0.400±0.033	0.195±0.009	0.395±0.012	0.241±0.007	0.378±0.013	0.340±0.010	0.319±0.007	0.226±0.010	0.332±0.011	0.282±0.008	0.352±0.010	0.334±0.009
	specific	0.387±0.010	0.260±0.009	0.408±0.005	0.290±0.003	0.403±0.008	0.326±0.009	0.325±0.003	0.242±0.008	0.343±0.005	0.272±0.009	0.349±0.006	0.323±0.006
BioMistral 7b 1000MIMIC FT	general	0.419±0.014	0.215±0.011	0.386±0.009	0.256±0.010	0.367±0.010	0.343±0.011	0.310±0.008	0.296±0.018	0.337±0.011	0.333±0.016	0.351±0.012	0.344±0.010
	specific	0.406±0.019	0.190±0.013	0.406±0.015	0.232±0.012	0.404±0.010	0.306±0.008	0.257±0.007	0.190±0.010	0.281±0.007	0.231±0.011	0.314±0.010	0.257±0.004

Table 8: Assessing the Impact of Augmented Data Sizes (respectively trained on 10, 100, and 1000 MIMIC notes)on
Open-Source LLM Training: Precision (P) and Recall (R)

Prompt Structure

```
### Instruction:
You are a helpful assistant, an expert in the medical domain. Extract top 3 main diagnosis/symptoms or conditions
mentioned in the medical note. Following the diagnosis/symptoms or conditions, identify the medical tests related to it. If
there aren't any medical tests related to it, just start listing the next important diagnosis/symptoms or conditions. If there
are no additional diagnoses/symptoms or conditions that you can identify, just list the existing ones and finalize the output.
Don't write no symptoms, or any indication that there is no other diagnosis/symptoms or conditions. Do not modify or
abbreviate what is written in the notes. Just extract them as they are. Make sure the highest priority is assigned with a
smaller number. We give you an example, do follow as below. The format should be as follows:

1. key symptom or condition
1.1 medical test related to 1
1.2 medical test related to 1
2. key symptom or condition
2.1 medical test related to 2
3. key symptom or condition
3.1 medical test related to 3
3.2 medical test related to 3

### Context:
{context}

### Response:
```

Figure 6: Structured Prompt

```
### Instruction:
You are a helpful assistant, an expert in medical domain. Extract top 3 key terms mentioned in the medical note that are
important for the patient. If you think they are of same importance, they can have the same ranking. Do not write no
symptoms, or any indication that there is no other diagnosis/symptoms or conditions. Do not modify or abbreviate what is
written in the notes. Just extract them as they are. Make sure the highest priority is assigned with a smaller number. We
give you an example, do follow as below. The format should be as follows

1. key term
1. key term
2. key term
2. key term
3. key term
3. key term

### Context:
{context}

### Response:
```

Figure 7: General Prompt

Instruction : You are a helpful assistant, an expert in medical domain. Extract top 10 main diagnosis/symptoms or conditions mentioned in the medical note. Following the diagnosis/symptoms or conditions, identify the medical tests related to it. If there isn't any medical tests related to it, just start listing the next important diagnosis/symptoms or conditions. If there are no additional diagnosis/symptoms or conditions that you can identify, just list the existing ones and finalize the output. Don't write no symptoms, or any indication that there is no other diagnosis/symptoms or conditions. Do not modify or abbreviate what is written in the notes. Just extract them as they are. Make sure the highest priority is assigned with a smaller number. We give you an example, do follow as below. The format should be as follows

1. key symptom or condition
1.1 medical test related to
1.2 medical test related to
2. key symptom or condition
2.1 medical test related to
3. key symptom or condition
3.1 medical test related to
3.2 medical test related to
4. key symptom or condition
4.1 medical test related to
4.2 medical test related to
4.3 medical test related to
5. key symptom or condition
5.1 medical test related to
6. key symptom or condition
6.1 medical test related to

[2 examples from the gold label] dataset

Figure 8: Prompt used for querying GPT-3.5 Turbo for data augmentation