Original Paper

Enhancing Large Language Models for Identifying and Prioritizing Important Medical Jargons From Electronic Health Record Notes Using Data Augmentation: Comparative Study

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

[1]Miner School of Computer and Information Sciences, University of Massachusetts Lowell, Lowell, MA, United States

[2]Manning College of Information & Computer Sciences, University of Massachusetts Amherst, Amherst, MA, United States

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

[*]these authors contributed equally

Corresponding Author:
Won Seok Jang, MS
Miner School of Computer and Information Sciences
University of Massachusetts Lowell
1 University Avenue
Lowell, MA, 01854
United States
Phone: 1 9789344000
Email: WonSeok_Jang@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: This study aimed to evaluate 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 90 expert-annotated EHR notes. We tested various combinations of settings, including (1) general and structured prompts, (2) zero-shot and few-shot prompting, (3) fine-tuning, and (4) data augmentation. To enhance the extraction and prioritization capabilities of open-source models in low-resource settings, we applied data augmentation using GPT-4o 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 9995 and tested their performance. The effectiveness of the models was assessed using 10-fold cross-validation, providing a comprehensive evaluation across various settings. We report the F_1-score and mean reciprocal rank for performance evaluation using two different string matching algorithms (relaxed string matching and Jaccard Index). We also conducted an error analysis classifying the erroneous outputs from the models.

Results: Our results show that open-source models achieved the highest performance, particularly when using fine-tuning with a gold-standard dataset. Under Jaccard Index–based string matching, DeepSeek 8B set the benchmarks with an F_1-score of 0.431 (SD 0.046); similarly, BioMistral 7B showed a mean reciprocal rank of 0.577 (SD 0.109). However, under relaxed string matching, open-source models were unable to match the performance of closed-source models, even with data augmentation or fine-tuning. We analyzed our experiment from several perspectives. First, few-shot prompting did not show an advantage over zero-shot prompting in vanilla models. Second, when comparing general and structured prompts, we found that model performance could deviate largely based on prompting styles. Third, fine-tuning with a small gold-standard dataset improved performance. Finally, data augmentation yielded performance comparable to or even surpassing the fine-tuning strategy. However, it also underscored the importance of the quality of the augmented dataset.

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 AI 2026;5:e75561) doi: 10.2196/75561

KEYWORDS

LLMs; large language models; data augmentation; EHR; electronic health record; comprehension; patient education; patient engagement

Introduction

Background

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 [3-9]. Nevertheless, the benefits of accessing EHR notes can diminish if patients do not comprehend their content [10-15]. EHR notes are lengthy and filled with medical jargon [16-19], which can be difficult to comprehend for the average US adult, whose reading ability is around the seventh- to eighth-grade level [20-24]. Therefore, supportive technologies are needed to assist patients in understanding EHR content [25,26], focusing on linking medical terms to lay-friendly terms [27-29], consumer-oriented definitions [18], and educational materials [30]. Early studies have demonstrated that such interventions significantly enhance patient understanding [18,27]. However, initial methods primarily relied on frequency- and context-based approaches to identify unfamiliar terms and propose simpler synonyms [27-29]. 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 [18,31-33].

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 [27-29]. 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 (eg, 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 natural language processing approaches to identify and rank jargon terms from EHR notes based on their importance to patients. LLMs have demonstrated tremendous promise in biomedical natural language processing applications [41-52] 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) [53-55], rather than on prioritizing terms most relevant to patients, which is crucial for enhancing communication between patients and health care providers.

The key contributions of this study are as follows:

1. We 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 using a physician-annotated gold-standard dataset.
2. We leverage data augmentation with AI-generated medical jargon from Medical Information Mart for Intensive Care IV (MIMIC-IV) discharge summaries to address the challenges of training in low-resource settings. Under relaxed string matching, none of the methods surpassed the performance of closed-source models. In contrast, evaluation using the Jaccard Index showed that fine-tuning and data augmentation improved performance, exceeding that of closed-source models.
3. 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 recommendations for users based on our findings.

Related Work

Identifying jargon terms important to patients is part of the 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, laboratory tests, and more [56,57]. Studies such as [58-61] have introduced language models for BioNER tasks, while more recent studies [53-55,62,63] have explored the application of 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 pretrained language models, such as Bidirectional Encoder Representations from Transformers (BERT [64]), Robustly Optimized BERT Pretraining approach (RoBERTa [59]), BioClinicalBERT [65], and BioBERT [58], 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 [66] have investigated whether LLMs, such as ChatGPT [67], can outperform baseline pretrained language models (eg, MedJEx [32] and SciSpacy [37]) in extracting personalized medical jargon. Similarly, GAMedX [68], a medical data extractor using LLMs (Mistral 7B and Gemma 7B), uses chained prompts to navigate the complexities of specialized medical jargon. Other works [69-71] 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 [72-74]. 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 [75], and FIT [35], FOCUS [34] uses MetaMap [36] to extract medical jargon from EHR notes and uses 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 [76] or help generate patient-friendly after-visit summaries [77].

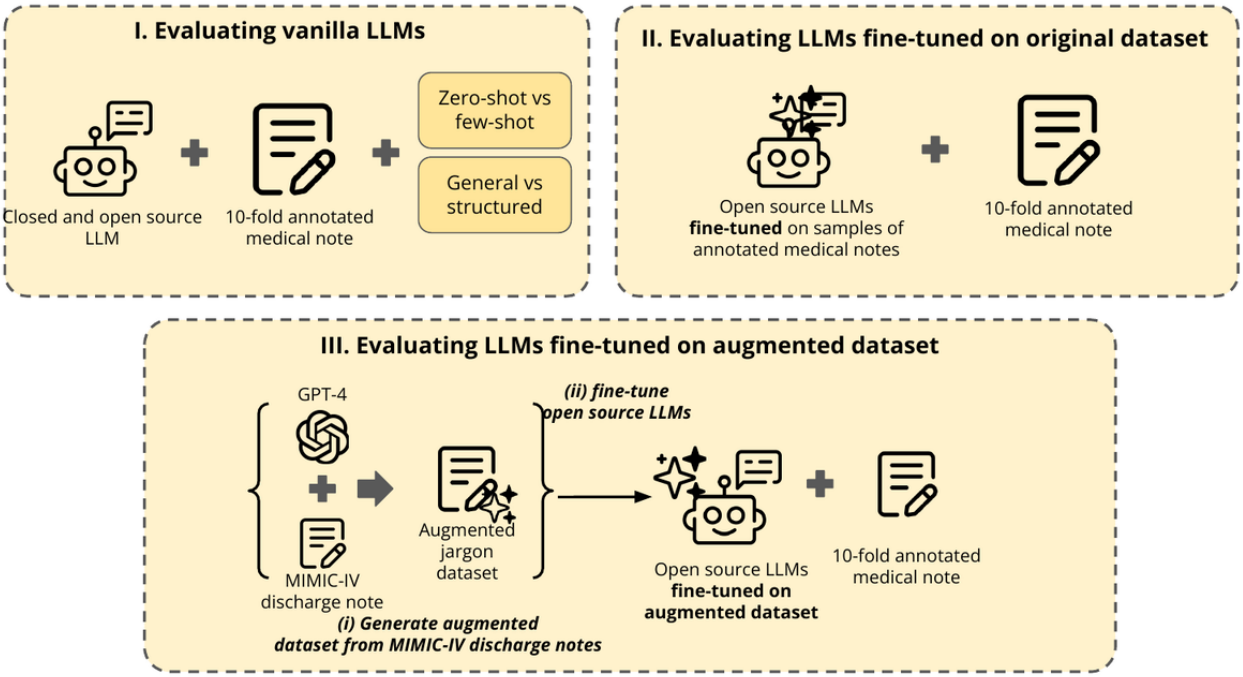
Methods

Overview

We evaluate the performance of the LLMs in three distinctive settings: (1) we assess the performance of closed- and open-source models by varying prompts and extraction tasks using the 10-fold annotated medical note (gold-standard dataset); (2) next, we benchmark the open-source models fine-tuned on portions of the gold-standard dataset; and (3) finally, we apply data augmentation, generating synthetic data from GPT-4o to fine-tune the open-source LLMs and evaluate them under the same varying settings. Our study finds that using data augmentation, the models can reach comparable or even superior performance in personalized medical jargon extraction tasks.

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 prompting styles, fine-tuning, and data augmentation to enhance model performance.

Figure 1. The evaluation workflow for closed- and open-source large language models (LLMs).



Data Source

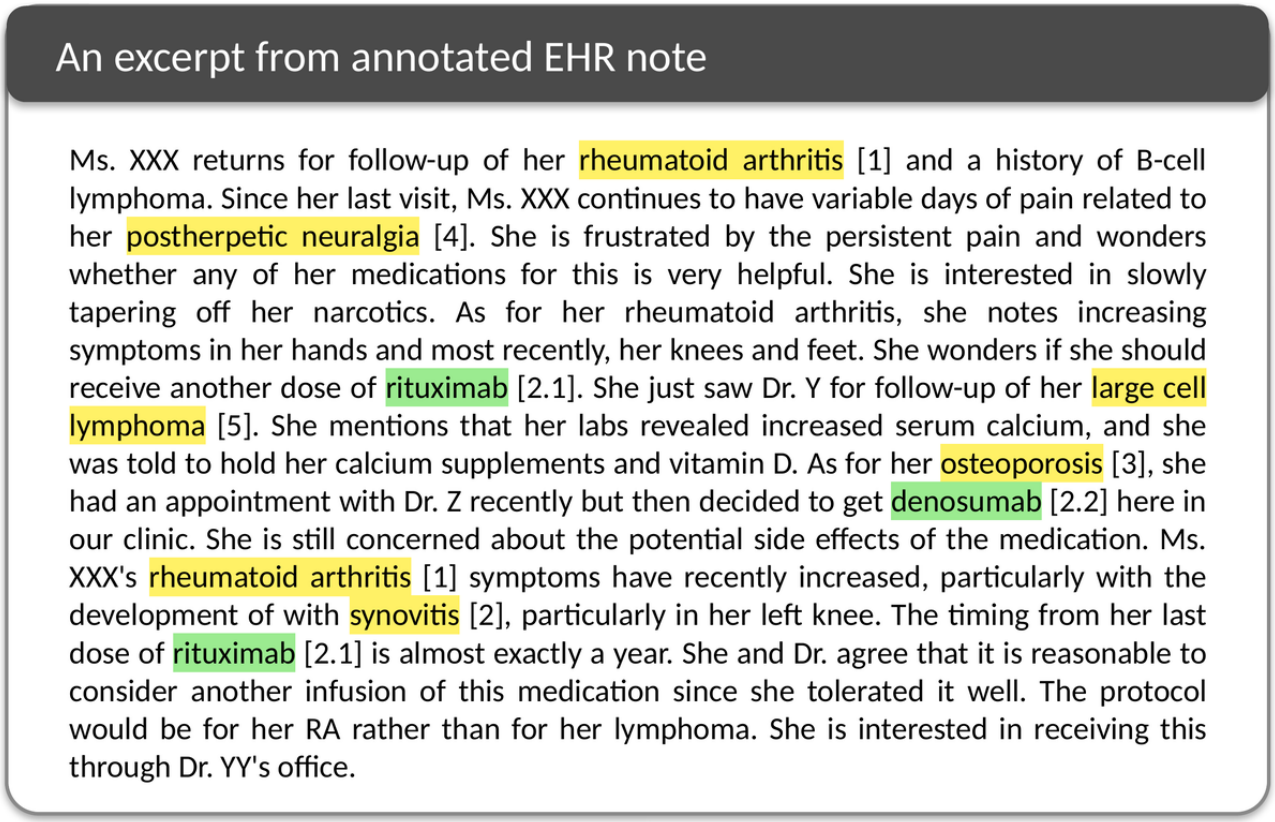
Our gold-standard dataset consists of 90 medical notes, each annotated by two physicians [34,35]. For each note, agreement from both physicians was used as the final annotation. The annotation agreement (micro average) on these notes was 0.51 Cohen Kappa [35]. This EHR note dataset comprises text reports across six medical categories: cancer, chronic obstructive pulmonary disease, diabetes, heart failure, hypertension, 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.

Table 1. Gold-standard dataset description. The dataset consists of 90 notes from patients diagnosed with cancer, chronic obstructive pulmonary disease, diabetes, hypertension, liver failure, and heart failure.

Main diagnosis	Note counts	Median (IQR) jargon counts
Cancer	17	8 (5-19)
COPD[a]	20	8 (5.8-13)
Diabetes	15	8 (6-12.5)
Hypertension	11	6 (4.5-8)
Liver failure	9	5 (4-10)
Heart failure	18	7 (5-10)

[a]COPD: chronic obstructive pulmonary disease.

Figure 2. A sample electronic health record note where physicians identified important medical terms. Diagnoses or conditions are highlighted in yellow, while medications, tests, and procedures associated with those diagnoses are marked in green, accompanied by their respective rankings. EHR: electronic health record.



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 Mistral 7B v0.1 [78] (Mistral 7B), BioMistral 7B [79], Llama3.1-8B [80] (Llama 3.1 8B), DeepSeek-R1-Llama-Distill-8B [81] (DeepSeek 8B). For proprietary models, we used 2 closed-source LLMs, which were from OpenAI [67]: GPT-5.2 and GPT-5-mini.

Zero-Shot Versus Few-Shot Prompts

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 dataset.

General and Structured Prompts

To evaluate the model's performance, we implemented two variations of prompts: general prompts and structured prompts, each designed to assess how effectively the model could extract relevant clinical information from medical notes. Multimedia Appendix 1 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 with them. In contrast, general prompts provide a more flexible and broader context. 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

To improve the performance of open-source LLMs (Mistral 7B, BioMistral 7B, Llama 3.1 8B, and DeepSeek 8B), we conducted low-rank adaptation [82] based on parameter-efficient fine-tuning. Low-rank adaptation 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 the low-rank dimension was set to 64. The learning rate was configured at $3e-4$, and the models were trained over 5 epochs to allow the models to converge effectively on the task-specific patterns present in the dataset.

Data Augmentation

Annotation by domain experts is expensive, and data augmentation using AI-generated data can help alleviate this challenge [83]. 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 [84]. We created an augmented dataset using the ICL technique, which was then used to refine the models (Multimedia Appendix 1). This augmented dataset was derived from discharge notes in the MIMIC-IV clinical database [85]. A subset of MIMIC-IV discharge notes was randomly selected, and GPT-4o from OpenAI [67] was used to process and rank key terms based on their importance for patient understanding. A dataset generated via GPT-4o was used because its outputs aligned more closely with the original annotations, exhibiting higher reliability during manual validation. The extraction process was guided by the examples, where two annotated notes from our gold-standard dataset were provided as examples to instruct the model on identifying and prioritizing terms.

After the raw generation, we filtered out terms that do not appear in the medical text using string matching. The resulting augmented notes were 9995 notes. Using only a few-shot technique and no other filtering mechanism, we anticipated a lot of noise being injected into the augmented dataset. To this end, we further conducted an analysis on the augmented dataset, collecting 100 random cases and comparing the synthetic annotations with an expert annotation. We asked a clinician to annotate the MIMIC-IV notes with the same instructions given to the GPT-4o and compared the agreement between them.

Exploring the Size of Augmented Dataset

To explore the effects of augmented dataset size, we progressively increased the dataset across four scales: 10, 100, 1000, and 9995 notes. By testing 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 for patients.

Baseline Models

We evaluated the performance of the open-source LLMs against several prominent baselines in the field of named entity recognition tasks: MedJEx [32] and BioClinical-ModernBERT [86]. For MedJEx, we did not fine-tune the model since it is already a trained model to extract medical jargon. Also, MedJEx does not output any confidence score; therefore, we could only report precision, recall, and F_1-score. For BioClinical-ModernBERT, we fine-tuned the model on our dataset and measured the performance. BioClinical-ModernBERT also outputs confidence scores. We used it as a proxy for ranking, assigning higher rankings for higher confidence scores. For the candidate entity extraction, we used MedspaCy [38] and QuickUMLS [39] to form inputs for the baseline models.

Evaluation Metrics

We used precision (Equation 1), recall (Equation 2), F_1-score (Equation 3), and mean reciprocal rank (MRR; Equation 4) as our metrics. Performance metrics, including precision, recall, and F_1-score, were first calculated at the individual note level (). To aggregate these into a single representative score for each of the 10 folds, we calculated the macroaverage across all notes within that fold. The final results reported are the mean and SD of these fold-level averages. Precision and recall gave insights into the model's ability to cover the annotated terms. The macro-F_1-score allowed us to evaluate the model's balanced performance. MRR metric (Equation 4), 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. Precision and recall are provided in the Multimedia Appendix 2. For MRR calculation, integer-valued ranks were used for both outputs from using the general prompt and the structured prompt. Because the structured prompt outputs used integer values in the MRR calculation, some results could receive tied ranks. To avoid inflating MRR, we used average ranks for tied results ($Adjusted\ Rank_i$). Specifically, when multiple terms shared the same rank, each term was assigned the average of the positions they occupied. This provides a more conservative evaluation by penalizing ambiguous predictions and rewarding models that rank the correct term more specifically. We used two different string matching approaches: (1) relaxed string matching, which checks the matching as true positives when it either perfectly matches or contains a gold-standard input [87,88], and (2) the Jaccard Index method that uses set-based similarity to identify matches, using Jaccard Distance to measure the gap between two strings. Here, we set the similarity threshold to >0.5. Any extracted term achieving a Jaccard similarity score greater than 0.5 against the gold-standard term was counted as a true positive for the subsequent precision and recall calculations. While the Jaccard Index method penalizes verbosity, relaxed string matching does not. This enables us to have a clear picture of the extraction performance and quality.

$$Precision_i = \frac{TP_i}{TP_i + FP_i} (1)$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i} (2)$$

$$F1_i = 2 \times \frac{Precision_i \cdot Recall_i}{Precsion_i + Recall_i} (3)$$

$$MRR = \frac{1}{N}\sum_{i=1}^{N} RR_i (4)$$

$$RR_i = \frac{1}{Adjusted\ Rank_i} = \left(\frac{1}{v}\sum_{j=k}^{k+v-1} j\right)^{-1} (5)$$

Here, RR_i denotes the reciprocal rank for each note *i*. *k* is the starting rank or the tied rank, and *v* is the count of ties (Equation 5). As shown in Equation 5, we compute the arithmetic mean of the ordinal positions occupied by tied terms to determine the effective rank. The final MRR is then derived by averaging these adjusted RR_i values across the total number of notes *N* in the dataset (Equation 4).

Error Classification

To further understand the nature of the model's behavior, we conducted an error analysis of the model's outputs. First, we classified the model's outputs that were erroneous. We used three types of errors: (1) spurious error (SE), where the model included terms that are not in the gold-standard dataset but are included in the medical note; (2) granularity error (GE), which reflects overly specific or overly broad extractions; and (3) misalignment error (ME), where the model simply did not understand the instruction and misbehaved. Using this taxonomy, we identified the erroneous behavior of the LLMs.

Experimental Details

All experiments were performed with two Nvidia A100 graphics processing units, each with 40 GB of memory, an Intel Xeon Gold 6230 CPU, and 192 GB of RAM. We used Python 3.9 and the Hugging Face transformers library [89] for our experiment. For the closed-source models, we used OpenAI's ChatGPT API [67]. To reduce the randomness of the experiment, we set the generation temperature to 0.1 for every model.

Ethical Considerations

The requirement for ethical approval and informed consent was waived by the institutional review board at the VA Bedford Health Care System. The experiments were performed in accordance with the Declaration of Helsinki. The clinical data used for data augmentation were obtained from the MIMIC-IV database, a publicly available, deidentified repository of EHRs from patients admitted to the emergency department or an intensive care unit at the Beth Israel Deaconess Medical Center in Boston, Massachusetts. Access to the database was granted following the completion of the required PhysioNet Credentialed Health Data Use Agreement 1.5.0.

Results

The highest baseline F_1-score was achieved by MedJEx (0.138, SD 0.031) using the Jaccard Index. For MRR, BioClinical-ModernBERT scored 0.087 (SD 0.146) and 0.098 (SD 0.151) using relaxed string matching and Jaccard Index, respectively. Despite being fine-tuned on the dataset, BioClinical-ModernBERT did not show the highest performance. As we expected, the conventional methods have limitations in extracting medical entities that are particularly important to the patient. Given that these models have a parameter size of 150M, considerably smaller than the generative models used for comparison, this outcome is perhaps unsurprising. Table 2 presents the performance of the baseline models, which include commonly used language models for medical information extraction.

Table 2. Performance of baseline models of F1-score and mean reciprocal rank using relaxed string matching (relaxed) and Jaccard Distance–based method (Jaccard).

Models	F_1-score (relaxed), mean (SD)	F_1-score (Jaccard), mean (SD)	MRR[a] (SD), relaxed	MRR (SD), Jaccard
MedJEx	0.120 (0.027)	0.138 (0.031)	—[b]	—
BioClinical-Modern BERT	0.076 (0.079)	0.087 (0.086)	0.087 (0.146)	0.098 (0.151)

[a]MRR: mean reciprocal rank.

[b]Not available.

The highest F_1-score in the zero-shot setting was 0.496 (SD 0.058), achieved by GPT-5.2. GPT-5.2 achieved the highest MRR in zero-shot prompts (0.578, SD 0.045) and (0.467, SD 0.117) with relaxed string matching and Jaccard Index. Notably, DeepSeek 8B and Llama 3.1 8B achieved higher Jaccard Index scores than GPT-5.2. As this metric is inversely affected by verbosity—additional non-overlapping tokens increase the union while leaving the intersection unchanged—GPT-5.2's more expansive outputs likely explain its lower performance. We also observed a substantial gap between relaxed string matching and Jaccard Index scores across models. For example, BioMistral 7B (Tables 3 and 4) exhibited a notable discrepancy in F_1-scores between these 2 metrics, indicating that vanilla models tended to produce verbose outputs. Tables 3 and 4 present the results from both closed- and open-source models using zero-shot and few-shot prompts, respectively. In Table 3, we compared the performance of different models using the top 5 and top 10 results for F_1-score and MRR across both closed- and open-source LLMs, providing their mean and SD. GPT-5.2 showed the highest performance of F_1-score and MRR using

relaxed string matching, but DeepSeek 8B also showed the highest F_1-score using the Jaccard Index. In Table 4, we compared the performance of different models using the top 5 and top 10 results for F_1-score and MRR across both closed- and open-source LLMs, providing their mean and SD. GPT-5.2 achieved the best F_1-score and MRR score using relaxed string matching. However, Mistral 7B also achieved a better F_1-score than GPT-5.2 using the Jaccard Index.

Table 3. Comparison between closed- and open-source vanilla models with zero-shot prompts.

Models	Prompt	Top 5				Top 10			
		F_1-score (relaxed), mean (SD)	F_1-score (Jaccard), mean (SD)	MRR[a] (SD), relaxed	MRR (SD), Jaccard	F_1-score (relaxed), mean (SD)	F_1-score (Jaccard), mean (SD)	MRR, SD (relaxed)	MRR, SD (Jaccard)
Closed-source LLMs [b]									
GPT-5.2									
	General	0.492 (0.068)[c]	0.307 (0.071)	0.578 (0.045)[c]	0.476 (0.124)[c]	0.496 (0.058)[c]	0.31 (0.077)	0.578 (0.045)[c]	0.467 (0.117)[c]
	Structured	0.332 (0.077)	0.206 (0.048)	0.513 (0.153)	0.359 (0.123)	0.336 (0.073)	0.211 (0.047)	0.513 (0.153)	0.359 (0.123)
GPT-5-mini									
	General	0.46 (0.073)	0.129 (0.048)	0.561 (0.156)	0.196 (0.11)	0.505 (0.063)	0.145 (0.045)	0.552 (0.151)	0.196 (0.11)
	Structured	0.387 (0.055)	0.12 (0.046)	0.575 (0.165)	0.24 (0.17)	0.387 (0.075)	0.127 (0.052)	0.566 (0.153)	0.24 (0.17)
Open-source LLMs									
Mistral 7B									
	General	0.372 (0.085)	0.323 (0.081)	0.451 (0.175)	0.453 (0.172)	0.401 (0.119)	0.345 (0.115)	0.432 (0.167)	0.444 (0.177)
	Structured	0.351 (0.083)	0.243 (0.056)	0.582 (0.108)	0.455 (0.112)	0.349 (0.081)	0.236 (0.055)	0.573 (0.106)	0.455 (0.112)
Llama 3.1 8B									
	General	0.347 (0.054)	0.333 (0.055)	0.426 (0.112)	0.392 (0.108)	0.368 (0.063)	0.35 (0.056)	0.422 (0.113)	0.392 (0.108)
	Structured	0.371 (0.056)	0.188 (0.051)	0.521 (0.14)	0.318 (0.112)	0.358 (0.059)	0.186 (0.055)	0.517 (0.139)	0.318 (0.112)
BioMistral 7B									
	General	0.21 (0.076)	0.017 (0.032)	0.353 (0.141)	0.033 (0.071)	0.211 (0.078)	0.02 (0.04)	0.353 (0.141)	0.033 (0.071)
	Structured	0.304 (0.121)	0.017 (0.018)	0.507 (0.152)	0.056 (0.056)	0.304 (0.123)	0.017 (0.018)	0.507 (0.152)	0.056 (0.056)
DeepSeek 8B									
	General	0.42 (0.077)	0.383 (0.053)[c]	0.416 (0.113)	0.398 (0.13)	0.443 (0.077)	0.409 (0.068)[c]	0.414 (0.113)	0.397 (0.13)
	Structured	0.326 (0.026)	0.211 (0.053)	0.467 (0.131)	0.4 (0.158)	0.328 (0.048)	0.218 (0.059)	0.467 (0.131)	0.4 (0.158)

[a]MRR: mean reciprocal rank.

[b]Highest score for each metric (column) among the compared models.

[c]LLM: large language model.

Table 4. Comparison between closed- and open-source vanilla models with few-shot prompts.

Models and prompts	Top 5				Top 10			
	F_1-score, mean (SD), relaxed	F_1-score, mean (SD), Jaccard	MRR[a] (SD), relaxed	MRR (SD), Jaccard	F_1-score, mean (SD), relaxed	F_1-score, mean (SD), Jaccard	MRR (SD), relaxed	MRR (SD), Jaccard
Closed-source LLMs[b]								
GPT-5.2								
General	0.475 (0.068)[c]	0.323 (0.06)	0.561 (0.098)[c]	0.536 (0.092)[c]	0.478 (0.07)[c]	0.326 (0.058)	0.561 (0.098)[c]	0.536 (0.092)[c]
Structured	0.354 (0.069)	0.233 (0.054)	0.536 (0.174)	0.387 (0.166)	0.364 (0.076)	0.237 (0.052)	0.536 (0.173)	0.387 (0.165)
GPT-5-mini								
General	0.475 (0.077)	0.243 (0.087)	0.543 (0.138)	0.4 (0.133)	0.492 (0.056)	0.245 (0.069)	0.543 (0.138)	0.4 (0.133)
Structured	0.348 (0.068)	0.161 (0.041)	0.508 (0.14)	0.321 (0.172)	0.347 (0.087)	0.163 (0.052)	0.495 (0.131)	0.323 (0.174)
Open-source LLMs								
Mistral 7B								
General	0.387 (0.046)	0.338 (0.039)	0.498 (0.141)	0.503 (0.11)	0.383 (0.05)	0.335 (0.045)	0.498 (0.141)	0.503 (0.11)
Structured	0.363 (0.067)	0.335 (0.056)	0.534 (0.104)	0.533 (0.101)	0.356 (0.073)	0.332 (0.065)	0.534 (0.104)	0.533 (0.101)
Llama 3.1 8B								
General	0.372 (0.078)	0.335 (0.093)	0.521 (0.145)	0.557 (0.126)	0.369 (0.073)	0.333 (0.096)	0.521 (0.145)	0.547 (0.132)
Structured	0.368 (0.116)	0.346 (0.111)[c]	0.442 (0.066)	0.502 (0.109)	0.374 (0.126)	0.353 (0.119)[c]	0.442 (0.066)	0.502 (0.109)
BioMistral 7B								
General	0.208 (0.063)	0.08 (0.036)	0.459 (0.115)	0.3 (0.132)	0.207 (0.067)	0.081 (0.033)	0.459 (0.115)	0.3 (0.132)
Structured	0.227 (0.085)	0.098 (0.042)	0.435 (0.072)	0.318 (0.136)	0.225 (0.084)	0.097 (0.041)	0.435 (0.072)	0.318 (0.136)
DeepSeek 8B								
General	0.344 (0.058)	0.335 (0.058)	0.494 (0.125)	0.485 (0.115)	0.344 (0.066)	0.334 (0.064)	0.485 (0.116)	0.476 (0.102)
Structured	0.327 (0.105)	0.299 (0.116)	0.48 (0.173)	0.466 (0.188)	0.332 (0.105)	0.304 (0.113)	0.48 (0.173)	0.466 (0.188)

[a]MRR: mean reciprocal rank.

[b]Highest score for each metric (column) among the compared models.

[c]LLM: large language model.

There was no clear advantage of using a few-shot prompt. The highest scores were achieved by using zero-shot prompts. Although variations still exist, most of the models (GPT-5.2, GPT-5-mini, Mistral 7B, and DeepSeek 8B) showed higher scores using a general prompt over a structured prompt. We hypothesize that each LLM has its own preferred prompts.

Fine-tuning open-source models resulted in better performance compared to vanilla models. Here, we unified the prompt style into a general prompt to evaluate the effectiveness of fine-tuning. The best-performing model was DeepSeek 8B, showing 0.424 (SD 0.052) in F_1-score (relaxed) and 0.431 (SD 0.046) in F_1-score (Jaccard) in the top 10 medical jargon extraction. For MRR, BioMistral 7B showed the best performance of 0.565 (SD 0.111) in MRR (relaxed) and 0.577 (SD 0.109) in MRR (Jaccard). Under relaxed string matching, fine-tuning did not outperform closed-source models such as GPT-5 (Tables 3-5). However, it yielded notable improvements in domain-specific task performance. Conversely, fine-tuned models exceeded closed-source models under the Jaccard Index metric, indicating that fine-tuning reduced verbosity while maintaining relevant content in the generated outputs. Table 5 demonstrates a reduced margin between the relaxed and Jaccard

metrics. In Table 5, we used zero-shot prompting with a general prompting style in the top 5 and top 10 medical jargon extraction. In the top 10 extraction task, DeepSeek 8B presented the highest F_1-score in both metrics (relaxed and Jaccard). However, BioMistral 7B showed the highest MRR score in both metrics.

Table 5. The average F1-score and mean reciprocal rank score of fine-tuned open-source models with 10-fold cross-validation.

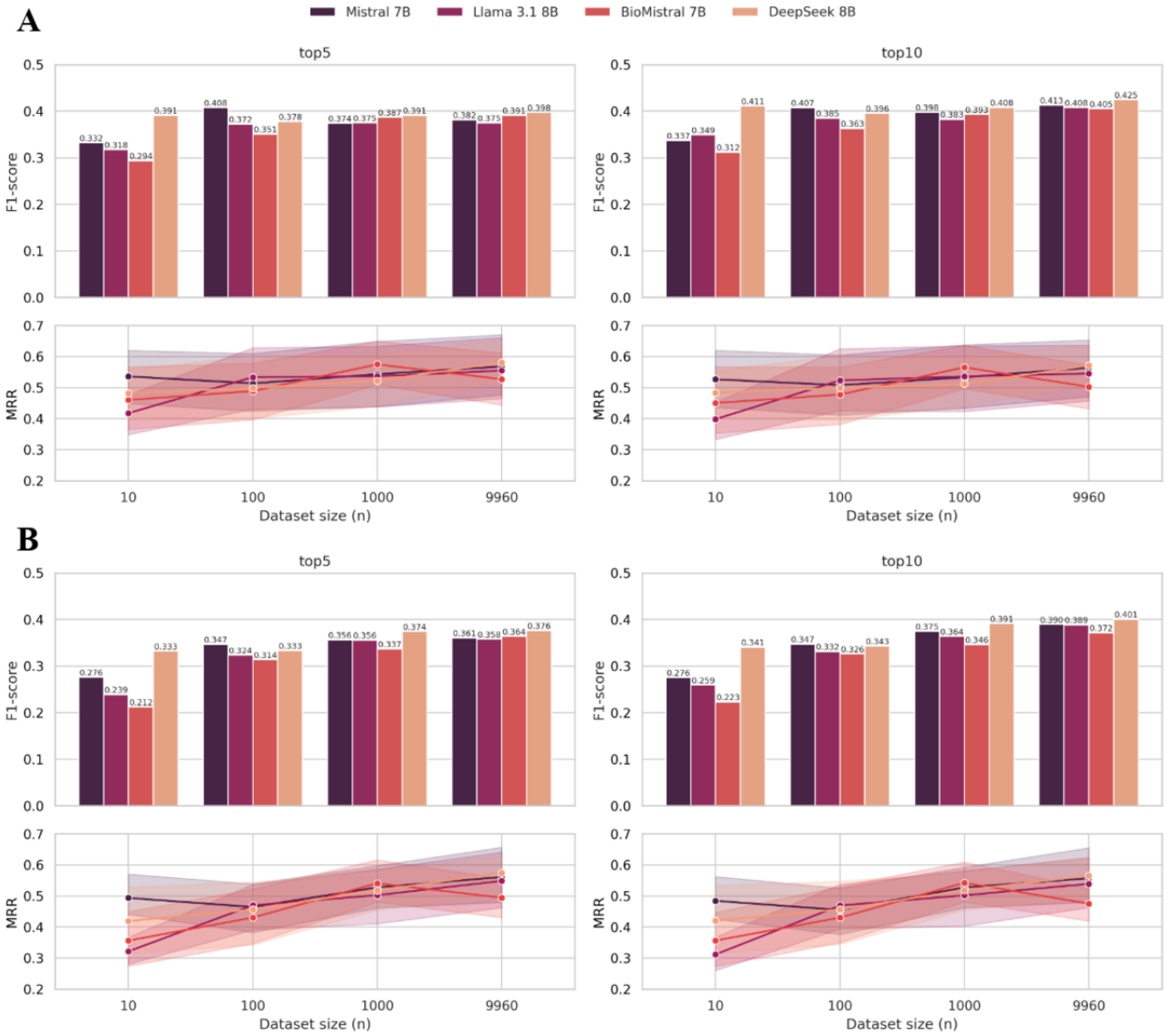
Models	Top 5				Top 10			
	F_1-score, mean (SD), relaxed	F_1-score, mean (SD), Jaccard	MRR[a] (SD), relaxed	MRR (SD), Jaccard	F_1-score, mean (SD), relaxed	F_1-score, mean (SD), Jaccard	MRR (SD), relaxed	MRR (SD), Jaccard
Mistral 7B	0.417 (0.093)	0.43 (0.074)[b]	0.525 (0.11)	0.573 (0.111)	0.416 (0.093)	0.428 (0.074)	0.525 (0.11)	0.573 (0.111)
Llama 3.1 8B	0.38 (0.065)	0.384 (0.058)	0.515 (0.078)	0.527 (0.103)	0.384 (0.07)	0.388 (0.064)	0.515 (0.078)	0.527 (0.103)
BioMistral 7B	0.374 (0.07)	0.379 (0.062)	0.565 (0.111)[b]	0.577 (0.109)[b]	0.377 (0.07)	0.381 (0.061)	0.565 (0.111)[b]	0.577 (0.109)[b]
DeepSeek 8B	0.42 (0.048)[b]	0.429 (0.043)	0.508 (0.085)	0.501 (0.075)	0.424 (0.052)[b]	0.431 (0.046)[b]	0.508 (0.085)	0.501 (0.075)

[a]MRR: mean reciprocal rank.

[b]Highest score for each metric (column) among the compared models.

The highest-performing model that used a data augmentation strategy was DeepSeek 8B, achieving an F_1-score (relaxed) above 0.425 and an F_1-score (Jaccard) of 0.401 in the top 10 medical jargon extractions (Figure 3; Tables S1 and S2 in Multimedia Appendix 2). The highest MRR score was achieved by DeepSeek 8B in the top 5 and top 10 medical jargon extractions using both metrics (relaxed and Jaccard). Performance patterns varied depending on the models and the metric. For example, under relaxed string matching, BioMistral 7B showed an increasing F_1-score trend, but MRR peaked at n=1000 before declining in both top 5 and top 10 extraction tasks. In contrast, Mistral 7B achieved peak performance at n=100 for top 5 extraction, while DeepSeek performed best at n=10 and n=9995. However, when evaluated using the Jaccard Index, all models exhibited consistently increasing F_1-scores as dataset size increased. These findings are discussed further below. Although the results failed to achieve higher scores than closed-source models under relaxed string matching, using the augmented dataset was found useful in some models, such as BioMistral 7B, Llama 3.1 8B, and DeepSeek 8B, but not in Mistral 7B. However, under the Jaccard Index, our results show that data augmentation could also outperform closed-source models.

Figure 3. Performance comparison of open-source large language models across varying scales of Medical Information Mart for Intensive Care IV–augmented data. Evaluation is based on relaxed string metrics (A) and Jaccard Distance (B), with training sizes ranging from 10 to 9995 samples. While individual model performance varies, a general trend of improvement is observed using the Jaccard Index (B) as the dataset size increases. Bar graphs represent F1-scores, while line graphs depict mean reciprocal rank. Both metrics indicate that DeepSeek 8B achieves the highest F1-score when trained on the largest dataset size and the highest mean reciprocal rank.



To further investigate error distributions, we evaluated the performance of the vanilla DeepSeek model against its fine-tuned and augmented counterparts (n=9995) using a representative test set from our 10-fold cross-validation. As illustrated in Table 6, the vanilla DeepSeek model demonstrated the highest frequency of MEs and GEs. While fine-tuning substantially reduced all error types, most notably SEs, which dropped to 46, the total error count was lowest for the fine-tuned model. Conversely, the data augmentation strategy did not yield similar improvements; it resulted in 80 SEs, surpassing the vanilla model, and only marginal reductions in GEs and MEs. Ultimately, the data augmentation approach failed to achieve the high-magnitude error reduction observed with standard fine-tuning.

Table 6. The error analysis table. All models show a high level of spurious error, but fine-tuning substantially reduces the error rate, showing the lowest total error counts among other methods.

Model versions	Spurious error, n	Granularity error, n	Misalignment error, n	Total error count, n
Vanilla model	76	7	37	120
Fine-tuned	46	1	33	80
Augmented dataset	80	2	33	115

Discussion

Zero-Shot vs Few-Shot

Our findings suggest that there are minimal differences between zero-shot and few-shot prompting. Although there are some cases where a few-shot prompt outperformed a zero-shot prompt, these results are not always consistent. Results differ by model, and it is hard to tell which prompting, zero-shot or few-shot, performs better. This is consistent with the findings reported by Chamieh et al [90], which found that few-shot prompting is not always more advantageous than zero-shot prompting (Figure 4).

Figure 4. Case study for extracting the top 10 important medical jargons from Mistral 7B in zero-shot and few-shot settings. Interestingly, few-shot prompting fails to extract terms that clinicians annotated as "gold," whereas zero-shot successfully conducts the task. COPD: chronic obstructive pulmonary disease; GERD: gastroesophageal reflux disease.

True labels	Mistral 7B zero-shot	Mistral 7B few-shot
1. Syncope	1. syncope	1. COPD
2. Cough-induced syncope	2. cough	1. sleep apnea
4. Arrhythmia	3. near syncope	1. coronary artery disease
4. Dysrhythmias	4. vertigo	1. GERD
7. Autonomic dysfunction	5. falls	1. hypertension
8. Benign positional vertigo	6. head trauma	1. osteoarthritis
8. Vertigo	7. fracture	1. carpal tunnel syndrome
	8. dysrhythmias	1. nephrolithiasis
	9. arrhythmia	1. degenerative disc disease
	10. EEG	1. depression
	11. MRI	1. erectile dysfunction
	12. brain	1. adenomatous polyps of the colon
	13. episodes	2. Losartan
	14. cough-induced syncope	…
	15. tonic-clonic activity	2. Iron sulfate
	…	2. Tylenol
	26. soft tissue attenuation	2. Oxycodone
	27. cardiac catheterization	2. Aspirin
	28. left main	2. Omeprazole
	29. left circumflex	2. Trileptal
	30. proximal right coron	2. Flomax

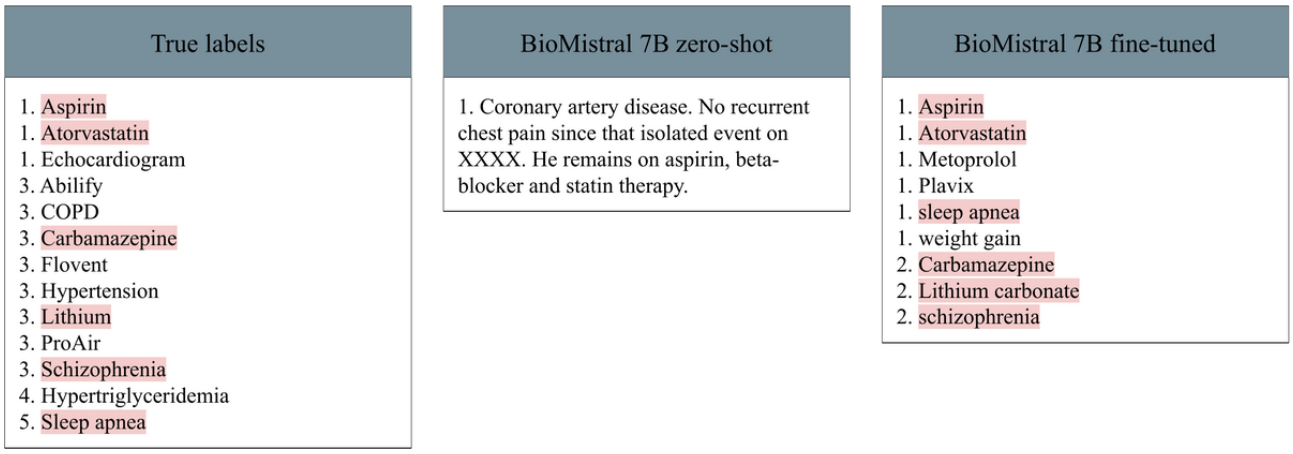
Prompting Styles

The effectiveness of prompts can vary across models, with certain prompt styles enhancing performance for specific models (Table 3). While differences between models were generally minimal, some models performed better with particular prompt styles. For example, in the Llama 3.1 8B model, structured prompts outperformed general prompts with few-shot prompting, whereas in Mistral 7B and in other models, general prompts showed improved performance over structured prompts regardless of string matching metrics. For both Llama 3.1 8B and Mistral 7B, structured prompts induced higher recall than precision. Since a structured format gives more specific instructions, this may cause the model to generate concise and precise results (Tables S1 and S2 in Multimedia Appendix 2). A plausible explanation is that structured prompts impose stronger output constraints, narrowing the model's generation space. Largely, our results align with prior research suggesting that tailored prompts can optimize a model's performance in specialized tasks [91]. Interestingly, in most cases, we found that F_1-score trends are aligned with MRR scores, suggesting that certain prompts helped the model reach its full potential. This shows that testing diverse prompt styles is essential in maximizing model performance, which is highly aligned with the current research results [92-95].

Fine-Tuning With Gold-Standard Data

Fine-tuning LLMs using domain-specific data proved effective in enhancing model performance, even surpassing the performance of closed-source models (GPT-5) under the Jaccard Index (Table 5). The size of the datasets used for fine-tuning was small but led to performance gains. Increasing the size of the fine-tuning dataset is likely to further improve the model's performance [96]. Figure 5 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. Although BioMistral 7B could extract some information with instructions, MEs were present in its output. After conducting fine-tuning, the model dramatically improved performance, reducing its errors as shown in Figure 5. This trend is further evidenced by the DeepSeek 8B results in Table 6, where the SE exhibits a substantial reduction. Furthermore, the verbosity of model outputs was significantly reduced after fine-tuning, as reflected in improved Jaccard Index scores and overall performance gains. 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 [58]. Moreover, instruction fine-tuning has been shown to improve the model's zero-shot performance [97].

Figure 5. Case study for extracting the top 10 important medical jargons from BioMistral 7B and fine-tuning BioMistral 7B on the gold-standard dataset. The fine-tuned model shows more robustness than vanilla models. The highlighted jargons are the ones that overlap with the expert-annotated labels, whereas BioMistral 7B shows misalignment erroneous output. After a few steps of fine-tuning, the model's behavior was significantly adjusted. Additionally, the verbosity of the output has decreased. COPD: chronic obstructive pulmonary disease.



Our results reveal that different string matching metrics yield divergent performance narratives. Closed-source models set the benchmark under relaxed string matching; however, fine-tuned models achieved superior F_1-scores and MRR when evaluated using the Jaccard Index. We contend that using a combination of string matching metrics enables a more nuanced assessment of model performance. Specifically, relaxed string matching prioritizes semantic preservation and the recognition of domain-specific terminology, while the Jaccard Index measures exact token-level overlap, thereby penalizing verbosity and rewarding concise outputs.

Analysis of Augmented Dataset

The data augmentation process relied solely on few-shot prompting, with string matching–based filtering as the only quality control measure. Consequently, the augmented dataset may be of lower quality than the gold-standard dataset. To evaluate this, we randomly sampled 100 instances from the augmented dataset and had a clinical expert manually annotate them. Agreement between the expert's annotations and the synthetic labels was measured using the Jaccard Index [98], and to measure the ranking agreement, MRR. The scores were 0.255 and 0.585, respectively. The terms extracted showed low Jaccard Distance, which likely contributed to lower performance. However, for MRR, the ranking showed comparable performance.

Fine-Tuning With Augmented Data

We explored the impact of data augmentation using various sizes of MIMIC-IV [85] discharge notes to enhance the performance of LLMs. Data augmentation simulated diverse scenarios within medical notes, enabling the LLM to generalize better across a broader range of note types. In accordance with fine-tuning, the performance gains from data augmentation were meaningful (Figure 3). In some models, we found that using augmented data was helpful in improving the performance (Llama 3.1 8B, BioMistral 7B, DeepSeek 8B) under relaxed string matching in the top10 extraction task. Furthermore, all models trained on the maximum dataset size (n=9995) demonstrated improved F_1-scores that exceeded those of closed-source models under the Jaccard index–based string matching, but not under relaxed string matching. These findings align with prior studies [97,99], 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 [100,101].

However, fine-tuning with an augmented dataset did not outperform fine-tuning with the gold-standard dataset for other models (Mistral 7B, BioMistral 7B, and DeepSeek 8B) when evaluated using Jaccard Index–based string matching (Table 5). Since we only used two examples for ICL, this may have affected the overall quality of the augmented dataset [102]. Nonetheless, this still leaves a question of why the LLMs' performance increased in the first place, considering the quality of the augmented dataset. Although data augmentation did not generate the most "accurate" synthetic dataset, it did not generate critical errors that severely harmed the model's performance. As a result, the augmented data provides weak but structured supervision that remains correlated with expert judgment. Consistent with prior findings in weakly supervised learning, such signals can still guide models toward meaningful patterns, particularly when aggregated across large-scale data. The observed performance improvements across models suggest that the augmented dataset captures clinically relevant structure despite instance-level disagreement, supporting the validity of our approach.

Impact of Augmented Dataset Size

The effect of augmented dataset size on model performance varied depending on the string matching metric. Under relaxed string matching, only BioMistral 7B demonstrated a consistent performance improvement with increasing dataset size (Figure 3A). Conversely, string matching using the Jaccard Index showed that all models achieved higher F_1-scores as the dataset size grew (Figure 3B), indicating metric-dependent performance trends. Our results partially align with findings from Yuan et al [103] and Kim et al [104], where increasing the size of synthetic datasets significantly enhanced model performance compared to models trained on smaller real-world datasets. For

other models such as DeepSeek 8B, the model's performance oscillated over dataset size in top 5 and top 10 extraction under relaxed string matching, although it ultimately showed the best performance at n=9995 regardless of the metrics.

Recommendations for Best Practices

Based on our findings, we categorize the risks of deploying LLMs into two distinct domains: operational risk and clinical risk. (1) Operational risk encompasses failures in prompt adherence or structural inconsistencies where the model's output deviates from the required format, potentially causing system-level failures. These risks are largely mitigable through rule-based validation and explicit format-checking mechanisms. (2) Clinical risk refers to the generation of factually incorrect information (hallucinations) that could directly compromise patient safety. Our results demonstrate that despite performance improvements across models, clinical risk remains a critical concern. Consequently, we argue that all LLM-generated outputs must undergo human-in-the-loop verification and be subjected to rigorous fact-checking protocols before clinical implementation.

Limitations

This study has several limitations. First, we tested only a limited selection of available closed- and open-source LLMs. Specifically, we only tested models with fewer than 10B parameters, which are relatively small open-source LLMs. Second, our gold-standard dataset, being based solely on physician annotations, captures only the clinical perspective, not considering the actual needs of patients. Third, despite the improvements achieved through fine-tuning and data augmentation, the performance of the models still falls short of human annotation. Fourth, the effectiveness of these methods should be validated in real-world settings, such as Hyper-DREAM, which tests the efficacy of the blood pressure management system [105].

Conclusions

Our study provides a comprehensive evaluation of closed- and open-source LLMs for identifying and prioritizing medical jargon within expert-annotated EHR notes. Strategic interventions, prompting, fine-tuning, and data augmentation notably enhanced the use of open-source models for domain-specific clinical tasks. Fine-tuning on a gold-standard dataset emerged as the most effective overall strategy, yielding the highest performance gains and a substantial reduction in error counts under Jaccard Index–based string matching. While the data augmentation strategy did not consistently outperform fine-tuning, it demonstrated measurable improvements in model-dependent scenarios. Nevertheless, our error analysis underscores that all evaluated models, regardless of architecture or optimization status, retain inherent degrees of both operational and clinical risk.

Acknowledgments

We greatly value the University of Massachusetts Biomedical Informatics Natural Language Processing (BioNLP) group's insightful feedback and thoughtful guidance. All the sections were revised through generative AI tools for checking grammatical errors and correcting awkward sentences. The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the US Department of Veterans Affairs or the US government.

Data Availability

The datasets generated and/or analyzed during the current study are not publicly available due to patient privacy and the terms of the data use agreement governing the source clinical notes. The source code will be released here [106].

Funding

The authors declared no financial support was received for this work.

Authors' Contributions

Conceptualization, methodology, software, formal analysis, investigation, data curation, visualization, and writing of the original draft: WSJ
Methodology, software, formal analysis, investigation, data curation, visualization, and writing of the original draft: SS
Project administration, writing of the original draft, and writing – review and editing: ZY
Methodology, validation, and writing – review and editing: ZY, HT, and SK
Conceptualization, supervision, project administration, funding acquisition, and writing – review and editing: HY

Conflicts of Interest

None declared.

Multimedia Appendix 1

Prompts used in data augmentation and prompting strategies.
[DOCX File , 260 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Additional results with mean (SD) of precision, recall.
[DOCX File , 66 KB-Multimedia Appendix 2]

References

1. Delbanco T, Walker J, Darer JD, Elmore JG, Feldman HJ, Leveille SG. Open notes: doctors and patients signing on. Ann Intern Med. Jul 20, 2010;153(2):121-125. [FREE Full text] [doi: 10.7326/0003-4819-153-2-201007200-00008] [Medline: 20643992]
2. Blue button. Office of the National Coordinator for Health Information Technology. 2024. URL: https://www.healthit.gov/patients-families/about-blue-button-movement [accessed 2026-06-12]
3. Delbanco T, Walker J, Bell SK, Darer JD, Elmore JG, Farag N, et al. Inviting patients to read their doctors' notes: a quasi-experimental study and a look ahead. Ann Intern Med. Oct 02, 2012;157(7):461-470. [FREE Full text] [doi: 10.7326/0003-4819-157-7-201210020-00002] [Medline: 23027317]
4. Gabay M. 21st Century Cures Act. Hosp Pharm. Apr 2017;52(4):264-265. [FREE Full text] [doi: 10.1310/hpj5204-264] [Medline: 28515504]
5. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc J. 2021;8(2):e188-e194. [FREE Full text] [doi: 10.7861/fhj.2021-0095] [Medline: 34286183]
6. Lye CT, Forman HP, Daniel JG, Krumholz HM. The 21st Century Cures Act and electronic health records one year later: will patients see the benefits? J Am Med Inform Assoc. 2018;25(9):1218-1220. [FREE Full text] [doi: 10.1093/jamia/ocy065] [Medline: 30184156]
7. Arvisais-Anhalt S, Lau M, Lehmann CU, Holmgren AJ, Medford RJ, Ramirez CM, et al. The 21st Century Cures Act and multiuser electronic health record access: potential pitfalls of information release. J Med Internet Res. Feb 17, 2022;24(2):e34085. [FREE Full text] [doi: 10.2196/34085] [Medline: 35175207]
8. Rodriguez JA, Clark CR, Bates DW. Digital health equity as a necessity in the 21st Century Cures Act era. JAMA. 2020;323(23):2381-2382. [doi: 10.1001/jama.2020.7858] [Medline: 32463421]
9. Nutbeam D. Artificial intelligence and health literacy—proceed with caution. Health Literacy and Communication Open. Oct 17, 2023;1(1):2263355. [doi: 10.1080/28355245.2023.2263355]
10. Root J, Oster NV, Jackson SL, Mejilla R, Walker J, Elmore JG. Characteristics of patients who report confusion after reading their primary care clinic notes online. Health Commun. 2016;31(6):778-781. [FREE Full text] [doi: 10.1080/10410236.2014.990078] [Medline: 26529325]
11. Kayastha N, Pollak KI, LeBlanc TW. Open oncology notes: a qualitative study of oncology patients' experiences reading their cancer care notes. JOP. Apr 2018;14(4):e251-e258. [doi: 10.1200/jop.2017.028605]
12. Kujala S, Hörhammer I, Väyrynen A, Holmroos M, Nättiaho-Rönnholm M, Hägglund M, et al. Patients' experiences of web-based access to electronic health records in Finland: cross-sectional survey. J Med Internet Res. Jun 06, 2022;24(6):e37438. [FREE Full text] [doi: 10.2196/37438] [Medline: 35666563]
13. Choudhry AJ, Baghdadi YM, Wagie AE, Habermann EB, Heller SF, Jenkins DH, et al. Readability of discharge summaries: with what level of information are we dismissing our patients? Am J Surg. Mar 2016;211(3):631-636. [FREE Full text] [doi: 10.1016/j.amjsurg.2015.12.005] [Medline: 26794665]
14. Khasawneh A, Kratzke I, Adapa K, Marks L, Mazur L. Effect of notes' access and complexity on OpenNotes' utility. Appl Clin Inform. Oct 2022;13(5):1015-1023. [FREE Full text] [doi: 10.1055/a-1942-6889] [Medline: 36104159]
15. Rahimian M, Warner JL, Salmi L, Rosenbloom ST, Davis RB, Joyce RM. 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. 2021;4(3):ooab051. [FREE Full text] [doi: 10.1093/jamiaopen/ooab051] [Medline: 34661067]
16. Zheng J, Yu H. Readability formulas and user perceptions of electronic health records difficulty: a corpus study. J Med Internet Res. Mar 02, 2017;19(3):e59. [FREE Full text] [doi: 10.2196/jmir.6962] [Medline: 28254738]
17. Zeng-Treitler Q, Kim H, Goryachev S, Keselman A, Slaughter L, Smith CA. Text characteristics of clinical reports and their implications for the readability of personal health records. Stud Health Technol Inform. 2007;129(Pt 2):1117-1121. [Medline: 17911889]
18. Polepalli RB, Houston T, Brandt C, Fang H, Yu H. Improving patients' electronic health record comprehension with NoteAid. Medinfo 2013 IOS Press. 2013:714-718. [doi: 10.3233/978-1-61499-289-9-714] [Medline: 23920650]
19. Sarzynski E, Hashmi H, Subramanian J, Fitzpatrick L, Polverento M, Simmons M, et al. Opportunities to improve clinical summaries for patients at hospital discharge. BMJ Qual Saf. May 2017;26(5):372-380. [doi: 10.1136/bmjqs-2015-005201] [Medline: 27154878]
20. Doak CC, Doak LG, Root JH. Teaching patients with low literacy skills. American Journal of Nursing. 1996;96(12):16M. [doi: 10.1097/00000446-199612000-00022]
21. Doak CC, Doak LG, Friedell GH, Meade CD. Improving comprehension for cancer patients with low literacy skills: strategies for clinicians. CA Cancer J Clin. 1998;48(3):151-162. [FREE Full text] [doi: 10.3322/canjclin.48.3.151] [Medline: 9594918]

22. Walsh TM, Volsko TA. Readability assessment of internet-based consumer health information. Respir Care. Oct 2008;53(10):1310-1315. [Medline: 18811992]
23. Eltorai AE, Han A, Truntzer J, Daniels AH. Readability of patient education materials on the American Orthopaedic Society for Sports Medicine website. Phys Sportsmed. 2014;42(4):125-130. [doi: 10.3810/psm.2014.11.2099] [Medline: 25419896]
24. Morony S, Flynn M, McCaffery KJ, Jansen J, Webster AC. Readability of written materials for CKD patients: a systematic review. Am J Kidney Dis. Jun 2015;65(6):842-850. [doi: 10.1053/j.ajkd.2014.11.025] [Medline: 25661679]
25. Johnson SB, Farach FJ, Pelphrey K, Rozenblit L. Data management in clinical research: synthesizing stakeholder perspectives. J Biomed Inform. Apr 2016;60:286-293. [FREE Full text] [doi: 10.1016/j.jbi.2016.02.014] [Medline: 26925516]
26. Morid MA, Fiszman M, Raja K, Jonnalagadda SR, Del Fiol G. Classification of clinically useful sentences in clinical evidence resources. J Biomed Inform. Apr 2016;60:14-22. [FREE Full text] [doi: 10.1016/j.jbi.2016.01.003] [Medline: 26774763]
27. Kandula S, Curtis D, Zeng-Treitler Q. A semantic and syntactic text simplification tool for health content. AMIA Annu Symp Proc. Nov 13, 2010;2010:366-370. [FREE Full text] [Medline: 21347002]
28. Zeng-Treitler Q, Goryachev S, Kim H, Keselman A, Rosendale D. Making texts in electronic health records comprehensible to consumers: a prototype translator. AMIA Annu Symp Proc. Oct 11, 2007;2007:846-850. [FREE Full text] [Medline: 18693956]
29. Abrahamsson E, Forni T, Skeppstedt M, Kvist M. Medical text simplification using synonym replacement: adapting assessment of word difficulty to a compounding language. 2014. Presented at: Proceedings of the 3rd Workshop on Predicting and Improving text Readability for Target Reader Populations (PITR); April 10, 2014:57-65; Gothenburg, Sweden. [doi: 10.3115/v1/w14-1207]
30. Zheng J, Yu H. Methods for linking EHR notes to education materials. Inf Retrieval J. Sep 03, 2015;19(1-2):174-188. [doi: 10.1007/s10791-015-9263-1] [Medline: 26306273]
31. Chen J, Druhl E, Polepalli Ramesh B, Houston TK, Brandt CA, Zulman DM, et al. A natural language processing system that links medical terms in electronic health record notes to lay definitions: system development using physician reviews. J Med Internet Res. Jan 22, 2018;20(1):e26. [FREE Full text] [doi: 10.2196/jmir.8669] [Medline: 29358159]
32. Kwon S, Yao Z, Jordan HS, Levy DA, Corner B, Yu H. MedJEx: a medical jargon extraction model with Wiki's hyperlink sspan and contextualized masked language model score. Proc Conf Empir Methods Nat Lang Process. Dec 2022;2022:11733-11751. [FREE Full text] [Medline: 37103473]
33. Leroy G, Endicott JE, Mouradi O, Kauchak D, Just ML. Improving perceived and actual text difficulty for health information consumers using semi-automated methods. AMIA Annu Symp Proc. 2012;2012:522-531. [FREE Full text] [Medline: 23304324]
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. JMIR Med Inform. Nov 30, 2016;4(4):e40. [FREE Full text] [doi: 10.2196/medinform.6373] [Medline: 27903489]
35. Chen J, Yu H. Unsupervised ensemble ranking of terms in electronic health record notes based on their importance to patients. J Biomed Inform. Apr 2017;68:121-131. [FREE Full text] [doi: 10.1016/j.jbi.2017.02.016] [Medline: 28267590]
36. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001:17-21. [FREE Full text] [Medline: 11825149]
37. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: fast and robust models for biomedical natural language processing. In: Association for Computational Linguistics. 2019. Presented at: Proceedings of the 18th BioNLP Workshop and Shared Task; August 1, 2019:190207669; Florence, Italy. [doi: 10.18653/v1/w19-5034]
38. Eyre H, Chapman AB, Peterson KS, Shi J, Alba PR, Jones MM, et al. Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. AMIA Annu Symp Proc. 2021;2021:438-447. [FREE Full text] [Medline: 35308962]
39. Soldaini L, Goharian N. QuickUMLS: a fast, unsupervised approach for medical concept extraction. 2016. Presented at: SIGIR MedIR Workshop; July 21, 2016:1-4; Pisa, Italy. URL: http://medir2016.imag.fr/data/MEDIR_2016_paper_16.pdf
40. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. Jan 01, 2004;32(Database issue):D267-D270. [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
41. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. Aug 2023;620(7972):172-180. [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
42. Tian S, Jin Q, Yeganova L, Lai PT, Zhu Q, Chen X, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Brief Bioinform. Nov 22, 2023;25(1):bbad493. [FREE Full text] [doi: 10.1093/bib/bbad493] [Medline: 38168838]
43. Tu T, Palepu A, Schaekermann M, Saab K, Freyberg J, Tanno R, et al. Towards conversational diagnostic AI. ArXiv. Preprint posted online on January 11, 2024. Feb 22, 2024;1(3):46. [doi: 10.1056/AIoa2300138]
44. McDuff D, Schaekermann M, Tu T, Palepu A, Wang A, Garrison J, et al. Towards accurate differential diagnosis with large language models. Nature. Jun 2025;642(8067):451-457. [FREE Full text] [doi: 10.1038/s41586-025-08869-4] [Medline: 40205049]

45. Wu C, Lin W, Zhang X, Zhang Y, Xie W, Wang Y. PMC-LLaMA: toward building open-source language models for medicine. J Am Med Inform Assoc. Sep 01, 2024;31(9):1833-1843. [doi: 10.1093/jamia/ocae045] [Medline: 38613821]
46. Chen Z, Cano AH, Romanou A, Bonnet A, Matoba K, Salvi F, et al. Meditron-70b: scaling medical pretraining for large language models. ArXiv. Preprint posted online on November 27, 2023. 2023. [FREE Full text] [doi: 10.48550/arXiv.2311.16079]
47. Tran H, Yang Z, Yao Z, Yu H. BioInstruct: instruction tuning of large language models for biomedical natural language processing. J Am Med Inform Assoc. Sep 01, 2024;31(9):1821-1832. [FREE Full text] [doi: 10.1093/jamia/ocae122] [Medline: 38833265]
48. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv. 2023. [doi: 10.5260/chara.21.2.8]
49. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. Feb 2023;2(2):e0000198. [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
50. Google Research, Google DeepMind. Advancing multimodal medical capabilities of Gemini. ArXiv. Preprint posted online on May 6, 2024. 2024:240503162. [FREE Full text] [doi: 10.48550/arXiv.2405.03162]
51. Yang Z, Yao Z, Tasmin M, Vashisht P, Jang WS, Wang B, et al. Performance of multimodal GPT-4V on USMLE with image: potential for imaging diagnostic support with explanations. medRxiv. Preprint posted online on November 15, 2023. 2023:10. [doi: 10.1101/2023.10.26.23297629]
52. Yao Z, Zhang Z, Tang C, Bian X, Zhao Y, Yang Z, et al. Medqa-cs: benchmarking large language models clinical skills using an AI-SCE framework. ArXiv. Preprint posted online on January 18, 2026. 2024:74. [FREE Full text]
53. Hu Y, Chen Q, Du J, Peng X, Keloth V, Zuo X, et al. Improving large language models for clinical named entity recognition via prompt engineering. J Am Med Inform Assoc. Sep 01, 2024;31(9):1812-1820. [FREE Full text] [doi: 10.1093/jamia/ocad259] [Medline: 38281112]
54. Monajatipoor M, Yang J, Stremmel J, Emami M, Mohaghegh F, Rouhsedaghat M. LLMs in biomedicine: a study on clinical named entity recognition. ArXiv. Preprint posted online on July 11, 2024. 2024. [FREE Full text] [doi: 10.48550/arXiv.2404.07376]
55. Hu D, Liu B, Zhu X, Lu X, Wu N. Zero-shot information extraction from radiological reports using ChatGPT. Int J Med Inform. Mar 2024;183:105321. [doi: 10.1016/j.ijmedinf.2023.105321] [Medline: 38157785]
56. Liu S, Wang A, Xiu X, Zhong M, Wu S. Evaluating medical entity recognition in health care: entity model quantitative study. JMIR Med Inform. 2024;12(1):e59782. [FREE Full text] [doi: 10.2196/59782] [Medline: 39419501]
57. Bose P, Srinivasan S, Sleeman WC, Palta J, Kapoor R, Ghosh P. A survey on recent named entity recognition and relationship extraction techniques on clinical texts. Applied Sciences. 2021;11(18):8319. [doi: 10.3390/app11188319]
58. Lee J, Yoon W, Kim S, Kim D, Kim S, So C, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. Feb 15, 2020;36(4):1234-1240. [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
59. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D. Roberta: A robustly optimized bert pretraining approach. ArXiv. Preprint posted online on July 26, 2019. 2019. [FREE Full text]
60. Yao Z, Cao Y, Yang Z, Deshpande V, Yu H. Extracting biomedical factual knowledge using pretrained language model and electronic health record context. AMIA Annu Symp Proc. 2022;2022:1188-1197. [FREE Full text] [Medline: 37128373]
61. Yao Z, Cao Y, Yang Z, Yu H. Context variance evaluation of pretrained language models for prompt-based biomedical knowledge probing. AMIA Jt Summits Transl Sci Proc. 2023;2023:592-601. [FREE Full text] [Medline: 37350903]
62. Gutierrez BJ, McNeal N, Washington C, Chen Y, Li L, Sun H, et al. Thinking about GPT-3 in-context learning for biomedical IE' think again. Association for Computational Linguistics. 2022:4497-4512. [doi: 10.18653/v1/2022.findings-emnlp.329]
63. Moradi M, Blagec K, Haberl F, Samwald M. Gpt-3 models are poor few-shot learners in the biomedical domain. ArXiv. Preprint posted online on September 6, 2021. 2021:210902555. [FREE Full text]
64. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. ArXiv. Preprint posted online on October 11, 2018. 2018. [FREE Full text]
65. Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings. ArXiv. Preprint posted online on April 6, 2019. Apr 6, 2019:1-7. [FREE Full text] [doi: 10.18653/v1/w19-1909]
66. Lim JH, Kwon S, Yao Z, Lalor JP, Yu H. Large language model-based role-playing for personalized medical jargon extraction. ArXiv. Preprint posted online on August 10, 2024. 2024. [FREE Full text]
67. OpenAI. URL: https://openai.com/ [accessed 2026-06-12]
68. Ghali MK, Farrag A, Sakai HE, Baz H, Jin Y, Lam S. Gamedx: generative AI-based medical entity data extractor using large language models. ArXiv. Preprint posted online on May 31, 2024. 2024:240520585. [doi: 10.2139/ssrn.5063216]
69. Butler JJ, Harrington MC, Tong Y, Rosenbaum AJ, Samsonov AP, Walls RJ, et al. From jargon to clarity: improving the readability of foot and ankle radiology reports with an artificial intelligence large language model. Foot Ankle Surg. Jun 2024;30(4):331-337. [doi: 10.1016/j.fas.2024.01.008] [Medline: 38336501]

70. Mannhardt N, Bondi-Kelly E, Lam B, O'Connell C, Mozannar H, Asiedu M, et al. Impact of large language model assistance on patients reading clinical notes: a mixed-methods study. ArXiv. Preprint posted online on January 17, 2024. 2024:240109637. [FREE Full text]
71. Lu J, Li J, Wallace BC, He Y, Pergola G. Napss: paragraph-level medical text simplification via narrative prompting and sentence-matching summarization. ArXiv. Preprint posted online on February 11, 2023. 2023:230205574. [doi: 10.18653/v1/2023.findings-eacl.80]
72. Speier W, Ong MK, Arnold CW. Using phrases and document metadata to improve topic modeling of clinical reports. J Biomed Inform. Jun 2016;61:260-266. [FREE Full text] [doi: 10.1016/j.jbi.2016.04.005] [Medline: 27109931]
73. Wen Z, Nair P, Deng CY, Lu XH, Moseley E, George N, et al. Mining heterogeneous clinical notes by multi-modal latent topic model. PLoS One. 2021;16(4):e0249622. [FREE Full text] [doi: 10.1371/journal.pone.0249622] [Medline: 33831055]
74. Sun S, Zack T, Williams CYK, Sushil M, Butte AJ. Topic modeling on clinical social work notes for exploring social determinants of health factors. JAMIA Open. Apr 2024;7(1):ooad112. [FREE Full text] [doi: 10.1093/jamiaopen/ooad112] [Medline: 38223407]
75. Chen J, Jagannatha AN, Fodeh SJ, Yu H. Ranking medical terms to support expansion of lay language resources for patient comprehension of electronic health record notes: adapted distant supervision approach. JMIR Med Inform. Oct 31, 2017;5(4):e42. [FREE Full text] [doi: 10.2196/medinform.8531] [Medline: 29089288]
76. Yao Z, Kantu NS, Wei G, Tran H, Duan Z, Kwon S, et al. README: Bridging medical jargon and lay understanding for patient education through data-centric NLP. Find ACL EMNLP. Nov 2024;2024:12609-12629. [doi: 10.18653/v1/2024.findings-emnlp.737] [Medline: 41710550]
77. Cai P, Liu F, Bajracharya A, Sills J, Kapoor A, Liu W, et al. Generation of patient after-visit summaries to support physicians. International Committee on Computational Linguistics; 2022. Presented at: Proceedings of the 29th International Conference on Computational Linguistics (COLING); October 12-17, 2022:6234-6247; Gyeongju, Republic of Korea. URL: https://aclanthology.org/2022.coling-1.544
78. Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, Casas de las D, et al. Mistral 7B. ArXiv. Preprint posted online on October 10, 2023. 2023. [FREE Full text] [doi: 10.48550/arXiv.2310.06825]
79. Labrak Y, Bazoge A, Morin E, Gourraud PA, Rouvier M, Dufour R. BioMistral: a collection of open-source pretrained large language models for medical domains. ArXiv. Preprint posted online on February 15, 2024. 2024. [FREE Full text] [doi: 10.18653/v1/2024.findings-acl.348]
80. Grattafiori A, Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, et al. The Llama 3 herd of models. ArXiv. Preprint posted online on July 31, 2024. 2024. [FREE Full text]
81. Guo D, Yang D, Zhang H, Song J, Wang P, Zhu Q, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature. Sep 2025;645(8081):633-638. [doi: 10.1038/s41586-025-09422-z] [Medline: 40962978]
82. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: low-rank adaptation of large language models. ArXiv. Preprint posted online on June 17, 2021. 2021. [FREE Full text] [doi: 10.5260/chara.21.2.8]
83. Li R, Wang X, Yu H. Two directions for clinical data generation with large language models: data-to-label and label-to-data. Proc Conf Empir Methods Nat Lang Process. 2023;2023:7129-7143. [FREE Full text] [doi: 10.18653/v1/2023.findings-emnlp.474] [Medline: 38213944]
84. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P. Language models are few-shot learners. 2020. Presented at: Proceedings of the 34th International Conference on Neural Information Processing Systems; December 6-12, 2020:200514165; Vancouver, BC. URL: https://papers.nips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf
85. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. Jan 03, 2023;10(1):1. [FREE Full text] [doi: 10.1038/s41597-022-01899-x] [Medline: 36596836]
86. Sounack T, Davis J, Durieux B, Chaffin A, Pollard T, Lehman E, et al. BioClinical ModernBERT: a state-of-the-art long-context encoder for biomedical and clinical NLP. ArXiv. Preprint posted online on June 12, 2025. 2025. [doi: 10.48550/arXiv.2506.10896]
87. Turney PD. Learning algorithms for keyphrase extraction. Information Retrieval. May 2000;2(4):303-336. [doi: 10.1023/a:1009976227802]
88. Zesch T, Gurevych I, Angelova G, Mitkov R. Approximate matching for evaluating keyphrase extraction. Association for Computational Linguistics; 2009. Presented at: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-2009); June 18, 2024:484-489; Borovets, Bulgaria. URL: https://aclanthology.org/R09-1086.pdf
89. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. HuggingFace's transformers: state-of-the-art natural language processing. ArXiv. Preprint posted online on October 9, 2019. 2019. [FREE Full text] [doi: 10.18653/v1/2020.emnlp-demos.6]
90. Chamieh I, Zesch T, Giebermann K. LLMs in short answer scoring: limitations and promise of zero-shot and few-shot approaches. Association for Computational Linguistics; 2024. Presented at: Proceedings of the 19th Workshop on Innovative

Use of NLP for Building Educational Applications (BEA 2024); June 20, 2024:309-315; Mexico City, Mexico. URL: https://aclanthology.org/2024.bea-1.25/ [doi: 10.18653/v1/2025.bea-1.0]
91. Liu J, Shen D, Zhang Y, Dolan B, Carin L, Chen W. What makes good in-context examples for GPT-3? ArXiv. Preprint posted online on January 17, 2021. 2021. [doi: 10.48550/arXiv.2101.06804]
92. Issaiy M, Ghanaati H, Kolahi S, Shakiba M, Jalali AH, Zarei D, et al. Methodological insights into ChatGPT's screening performance in systematic reviews. BMC Med Res Methodol. Mar 27, 2024;24(1):78. [FREE Full text] [doi: 10.1186/s12874-024-02203-8] [Medline: 38539117]
93. Krag CH, Balschmidt T, Bruun F, Brejnebøl M, Xu JJ, Boesen M, et al. Large language models for abstract screening in systematic- and scoping reviews: a diagnostic test accuracy study. medRxiv. Preprint posted online on October 2, 2024. 2024. [doi: 10.1101/2024.10.01.24314702]
94. Guo E, Gupta M, Deng J, Park YJ, Paget M, Naugler C. Automated paper screening for clinical reviews using large language models: data analysis study. J Med Internet Res. Jan 12, 2024;26:e48996. [FREE Full text] [doi: 10.2196/48996] [Medline: 38214966]
95. Homiar A, Thomas J, Ostinelli EG, Kennett J, Friedrich C, Cuijpers P, et al. Development and evaluation of prompts for a large language model to screen titles and abstracts in a living systematic review. BMJ Ment Health. Jul 22, 2025;28(1). [FREE Full text] [doi: 10.1136/bmjment-2025-301762] [Medline: 40701625]
96. 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. ArXiv. Preprint posted online on September 5, 2024. 2024. [doi: 10.48550/arXiv.2409.03454]
97. Wei J, Bosma M, Zhao VY, Guu K, Yu AW, Lester B, et al. Finetuned language models are zero-shot learners. ArXiv. Preprint posted online on September 3, 2021. 2022:1-42. [FREE Full text] [doi: 10.48550/arXiv.2109.01652]
98. Jaccard P. The distribution of the flora in the Alpine zone. New Phytologist. May 05, 2006;11(2):37-50. [doi: 10.1111/j.1469-8137.1912.tb05611.x]
99. Tang R, Han X, Jiang X, Hu X. Does synthetic data generation of LLMs help clinical text mining? ArXiv. Preprint posted on March 8, 2023. 2023. [FREE Full text] [doi: 10.5260/chara.21.2.8]
100. Wei J, Zou K. EDA: easy data augmentation techniques for boosting performance on text classification tasks. ArXiv. Prepint posted online on August 25, 2019. 2019. [doi: 10.18653/v1/d19-1670]
101. Chen J, Tam D, Raffel C, Bansal M, Yang D. An empirical survey of data augmentation for limited data learning in NLP. ArXiv. Preprint posted online on June 14, 2021. 2021. [FREE Full text] [doi: 10.1162/tacl_a_00542]
102. Iskander S, Cohen N, Karnin Z, Shapira O, Tolmach S. Quality matters: evaluating synthetic data for tool-using LLMs. Association for Computational Linguistics; 2024. Presented at: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing Miami; May 19, 2025:4958-4976; Florida, USA. [doi: 10.18653/v1/2024.emnlp-main.285]
103. Yuan J, Zhang J, Sun S, Torr P, Zhao B. Real-fake: effective training data synthesis through distribution matching. ArXiv. Preprint posted online on October 16, 2023. 2024:1-21. [FREE Full text] [doi: 10.48550/arXiv.2310.10402]
104. Kim S, Suk J, Yue X, Viswanathan V, Lee S, Wang Y, et al. Evaluating language models as synthetic data generators. ArXiv. Preprint posted online on September 1, 2025. 2024. [doi: 10.18653/v1/2025.acl-long.320]
105. Wang Y, Zhu T, Zhou T, Wu B, Tan W, Ma K. Hyper-DREAM, a multimodal digital transformation hypertension management platform integrating large language model and digital phenotyping: multicenter development and initial validation study. J Med Syst. Apr 02, 2025;49(1):42. [doi: 10.1007/s10916-025-02176-1] [Medline: 40172683]
106. Enhancing for identifying and prioritizing important medical jargons from electronic health record notes utilizing data augmentation : a pilot study. GitHub. URL: https://github.com/memy85/2024_medicalnote_annotation [accessed 2026-06-23]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
BioNER: biomedical named entity recognition
EHR: electronic health record
GE: granularity error
ICL: in-context learning
LLM: large language model
ME: misalignment error
MIMIC-IV: Medical Information Mart for Intensive Care IV
MRR: mean reciprocal rank
RoBERTa: Robustly Optimized BERT Pretraining approach
SE: spurious error
UMLS: Unified Medical Language System

Edited by A Coristine; submitted 06.Apr.2025; peer-reviewed by Y Wang, X Du, F Manion; comments to author 22.May.2025; accepted 24.Apr.2026; published 17.Jul.2026

Please cite as:
Jang WS, Sultana S, Yao Z, Tran H, Yang Z, Kwon S, Yu H
Enhancing Large Language Models for Identifying and Prioritizing Important Medical Jargons From Electronic Health Record Notes Using Data Augmentation: Comparative Study
JMIR AI 2026;5:e75561
URL: https://ai.jmir.org/2026/1/e75561
doi: 10.2196/75561
PMID: