Large Language Models Struggle in Token-Level Clinical Named Entity Recognition

Qiuhao Lu, Ph.D., Rui Li, Ph.D., Andrew Wen, M.S., Jinlian Wang, Ph.D., Liwei Wang, M.D., Ph.D., Hongfang Liu, Ph.D.
McWilliams School of Biomedical Informatics, University of Texas Health Science Center, Houston, TX, USA

Abstract

Large Language Models (LLMs) have revolutionized various sectors, including healthcare where they are employed in diverse applications. Their utility is particularly significant in the context of rare diseases, where data scarcity, complexity, and specificity pose considerable challenges. In the clinical domain, Named Entity Recognition (NER) stands out as an essential task and it plays a crucial role in extracting relevant information from clinical texts. Despite the promise of LLMs, current research mostly concentrates on document-level NER, identifying entities in a more general context across entire documents, without extracting their precise location. Additionally, efforts have been directed towards adapting ChatGPT for token-level NER. However, there is a significant research gap when it comes to employing token-level NER for clinical texts, especially with the use of local open-source LLMs. This study aims to bridge this gap by investigating the effectiveness of both proprietary and local LLMs in token-level clinical NER. Essentially, we delve into the capabilities of these models through a series of experiments involving zero-shot prompting, few-shot prompting, retrieval-augmented generation (RAG), and instruction-fine-tuning. Our exploration reveals the inherent challenges LLMs face in token-level NER, particularly in the context of rare diseases, and suggests possible improvements for their application in healthcare. This research contributes to narrowing a significant gap in healthcare informatics and offers insights that could lead to a more refined application of LLMs in the healthcare sector.

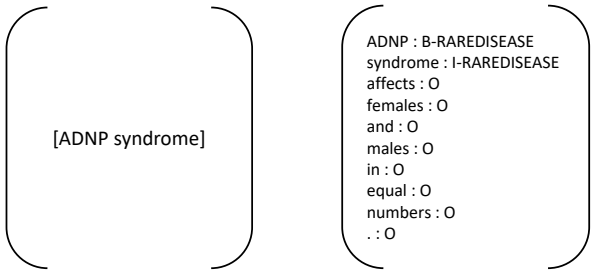
Introduction

Electronic Health Records (EHRs) are a key component in modern healthcare, encapsulating vast amounts of patient data, most notably in clinical notes. Their widespread adoption by healthcare providers in recent years has revolutionized the manner in which patients' visits and health information are recorded and managed[1], thereby elevating the quality of patient care. However, extracting pertinent information from these records presents a significant challenge due to the abundant, heterogeneous, and private nature of clinical texts. At the heart of addressing this challenge is Named Entity Recognition (NER), a fundamental task in natural language processing aimed at identifying and categorizing key information units in the text. In the clinical domain, the goal of the task is to identify all occurrences of specific clinically relevant named entities in the given unstructured clinical narrative[2].

There has been a surge of interest in developing clinical NER systems in the last few years. Early methods mostly rely on manually-crated rules and traditional machine learning techniques, such as MetaMap[3], KnowledgeMap[4], cTAKES[5], etc. With the trending of deep learning methods and the Transformer architecture[6], more researchers shift to building clinical NER systems and other NLP applications upon pre-trained language models such as BERT[7]. A typical solution is to insert a multilayer perception (MLP) on top of the language model and train the entire model via fine-tuning. For example, Alsentzer *et al.* fine-tune BioClinicalBERT on four i2b2 NER tasks[8]. Similarly, Sung *et al.* present BERN2, which uses Bio-LM[9] as the foundation model and achieves better performance via multitask learning[10].

A common theme among the aforementioned deep-learning-based methods is their data-hungry nature, which unfortunately poses significant challenges in the medical field, where issues of data scarcity, heterogeneity, and confidentiality are consistently prevalent. The situation is even worse for *rare diseases* due to their diversity, complexity, and specificity. Rare diseases refer to diseases or conditions that affect fewer than $200,000$ Americans by definition in the United States[11]. These conditions are often underrepresented in datasets, leading to a scarcity of information that deep learning models can learn from.

Recent advancements in computational linguistics have led to the emergence of Large Language Models (LLMs). These models have shown remarkable capabilities in various domains, including the clinical field[12]. Their strengths in



(a) document-level **(b)** token-level

Figure 1: Different NER outputs for the given text: *ADNP syndrome affects females and males in equal numbers* .

in-context learning, such as retrieval-augmented generation (RAG), present new opportunities for tackling the challenges of NER in clinical texts. Despite the promise of LLMs, most existing research in this area has either focused on document-level NER (i.e., aggregation NER), which involves identifying entities at a broader level in entire documents or paragraphs without predicting the exact span or location of these entities, or on adapting ChatGPT to this task. For example, Zhou *et al.* introduce UniversalNER, a distillation approach using mission-focused instruction-tuning to create efficient models that excel in NER. They distill ChatGPT into a more compact LLaMA-based[13] model UniversalNER and demonstrate superior document-level NER accuracy across diverse domains, including biomedicine[14]. Similarly, in their exploration of ChatGPT-3.5-turbo, Shry *et al.* focus on the extraction of rare diseases and their associated phenotypes via prompt engineering, also limiting their analysis to document-level output, which overlooks the precise span and location of these entities[15]. In contrast, the potential of LLMs, especially local open-source LLMs, in token-level NER (i.e., span-based NER) in clinical settings, particularly for rare diseases, remains relatively unexplored.

The difference between document-level and token-level NER is illustrated in Figure 1. This gap is significant as token-level NER offers the potential for more detailed and precise entity recognition, which is crucial in clinical applications. For instance, to interpret clinical texts like chemotherapy treatment records, in a scenario where a patient's record contains multiple chemotherapy sessions over several months, document-level NER may not efficiently distinguish between the different treatment phases and their respective impacts. In contrast, token-level NER can capture specific details. For example, consider a clinical text, "Patient experienced nausea after chemo session 1 in January and significant fatigue following chemo session 3 in March." While document-level NER might identify terms like "chemo", "nausea", and "fatigue", it fails to connect these conditions to specific treatment events. Token-level NER, however, can extract fine-grained details, such as linking "nausea" directly to "chemo session 1 in January", providing valuable context for understanding and tracking the patient's treatment response over time. This precision is vital for personalized patient care.

This study aims to explore the effectiveness of LLMs in token-level NER for rare diseases within clinical texts. By focusing on this specific application, we seek to understand the limitations and capabilities of LLMs in handling the challenges of token-level entity recognition in a domain where data is scarce and highly specialized. Essentially, we explore the performance of various state-of-the-art local open-source LLMs, including LLaMA-2[16], Meditron[17], Llama2-MedTuned[18], and UniversalNER[14], on the task of NER of rare diseases and their phenotypes. Additionally, we assess the performance of prominent models such as ChatGPT-3.5 and ChatGPT-4. Our investigation extends to examining their capabilities in in-context learning, evaluating performance enhancements through few-shot learning, retrieval-augmented generation (RAG), and fine-tuning. The experimental results reveal that without proper fine-tuning, local LLMs generally struggle with token-level NER, despite prior training on clinical texts. We observe that while few-shot learning can effectively improve the performance of most LLMs, the influence of RAG on the same task is relatively minimal. Notably, our findings indicate that a medically-adapted LLaMA-2-7b model, specifically Llama2-MedTuned[18], can outperform ChatGPT-4 on this task. This is particularly noteworthy as it achieves these results without having been trained specifically on the rare disease data used in the experiments, and this highlights the potential of fine-tuning local LLMs for specialized applications in clinical NER and other clinical NLP tasks.

Table 1: Statistics of the RareDis dataset.

Data Type	Training	Validation	Test	Total
Documents	729	104	208	1041
Sentences	6281	894	1735	8910
Disease	1328	187	392	1907
Rare Disease	3075	468	918	4461
Skin Rare Disease	442	42	143	627
Sign	3384	474	966	4824
Symptom	313	22	51	386

Related Work

Recently, researchers have released various LLMs dedicated to the medical field. LLaMA[13] and LLaMA-2[16] are one of the first and most popular open-source LLMs, laying the foundation for extensive research and applications. Meditron[17] is a suite of open-source medical LLMs pre-trained on the GAP-Replay corpus, including clinical guidelines, research papers, and general domain data, and surpass the performance of LLaMA-2[16], ChatGPT-3.5, and Flan-PaLM[19] on multiple medical reasoning benchmarks.

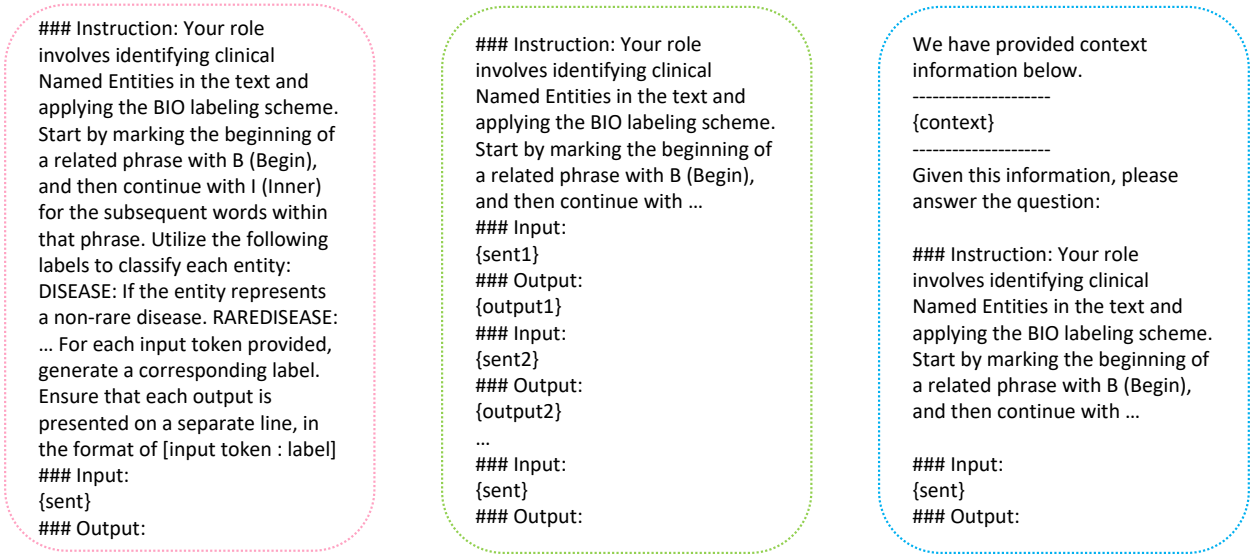
In parallel, considerable efforts have been devoted to adapting these LLMs to the task of NER through prompting or fine-tuning[13,14,15,20,21,22,23,24]. For example, Zhou *et al.* instruction-fine-tune the LLaMA model[13] using ChatGPT-generated synthetic data for NER from broad-coverage unlabeled web text[14]. Their UniversalNER model shows promising NER performance across multiple domains. However, they only generate document-level outputs, i.e., a list of extracted entities in the given text as shown in Figure 2a, without considering their exact span information, which limits their impact and usage in practical scenarios. Hu *et al.* explore the token-level NER capabilities of ChatGPT-3.5 and ChatGPT-4 on two clinical NER tasks by manually crafting specific prompts[20]. However, they limit their exploration to ChatGPT models and do not investigate local and open-source LLMs. Another study that aligns closely with our research is by Shry *et al.*, who focus on extracting rare diseases and their phenotypes at a document-level by prompting ChatGPT-3.5[15].

Unlike these studies, we aim to bridge the gap by exploring the capabilities of both local open-source LLMs and ChatGPT models in the specific context of token-level NER in clinical settings. This approach seeks to understand how these models perform in an area that is not only highly specialized but also characterized by its complexity and data scarcity, particularly in the study of rare diseases.

Methods

Task Overview Token-level named entity recognition (NER), also known as span-based NER, involves identifying and classifying named entities within a text into predefined categories such as diseases, symptoms, genes, etc. Formally, in a sentence with N tokens $X = [x_1, x_2, \ldots, x_N]$, an entity is defined as a contiguous span of tokens $e = [x_i, \ldots, x_j]$, where $0 \leq i \leq j \leq N$, and each is associated with a specific entity type. The core task is treated as a sequence labeling problem, where the goal is to assign a corresponding sequence of labels $Y = [y_1, y_2, \ldots, y_N]$ to the sentence X. In this study, the BIO (Beginning, Inside, Outside) schema is employed for labeling. According to this scheme, the first token of an entity of a certain `type` is tagged as `B-type`, any subsequent tokens within the same entity are tagged as `I-type`, and tokens not part of an entity are tagged as `O`. In this regard, the following eleven categories or labels are used in the experiments: `O`, `B-Disease`, `I-Disease`, `B-RareDisease`, `I-RareDisease`, `B-SkinRareDisease`, `I-SkinRareDisease`, `B-Symptom`, `I-Symptom`, `B-Sign`, and `I-Sign`.

In this study, we aim to understand the capabilities of various general and medical LLMs in token-level NER, which is underexplored in clinical settings. Similar to Shry *et al.*'s work[15], we choose rare diseases as our case study. Our focus is particularly on extracting information about rare diseases and their associated phenotypes, driven by two primary reasons: 1) Data related to rare diseases is scarce and highly specialized, presenting an ideal scenario to



(a) zero-shot **(b)** few-shot **(c)** RAG

Figure 2: Prompt templates for zero-shot, few-shot, and RAG-based inference.

leverage the strengths of LLMs; and 2) Patients with rare diseases are a relatively understudied group, necessitating increased research and advocacy efforts, as highlighted by Nguyen *et al.*[25] To this end, our investigation spans across the overall performance of both local open-source LLMs and ChatGPT models on this task. More specifically, we initially explore the zero-shot performance of these LLMs using manually-designed prompts. Secondly, to enhance the models' performance for this specific task, we employ in-context learning strategies, i.e., few-shot learning and retrieval-augmented generation (RAG). These approaches provide rich information to the prompts to facilitate a deeper understanding of the task by the LLMs. Finally, we instruction-fine-tune Llama2-MedTuned[18] on the rare disease dataset to assess the performance of LLMs in a fully supervised way. Our intention is to explore how well LLMs, when fine-tuned with direct supervision, can adapt to and accurately identify entities in complex rare disease data.

Dataset In our experiments, we use the RareDis-v1 dataset[26,27], a curated collection of texts from the National Organization for Rare Disorders (NORD) database[1]. This dataset specifically comprises selected sections from NORD articles, which have been manually annotated to identify five key entity types: `Disease`, `Rare Disease`, `Skin Rare Disease`, `Symptom`, and `Sign`. The statistics of the dataset are detailed in Table 1, where the first two rows show the number of documents and sentences, respectively. The last four rows are a breakdown of the specific clinical entities present in the dataset. It is important to distinguish between `Sign`, which are objectively observable indicators or test results suggesting a disease, and `Symptom`, which are subjective experiences reported by the patient[26].

Models We consider the following LLMs for our experiments:

- LLaMA-2-7b[16] is one of the first and most popular local open-source pre-trained LLMs released. Though it is not dedicated to the clinical domain, the model is often considered a critical benchmark in recent related studies due to its impact and generalizability. For consistency, we utilize the 7-billion-parameter version of all local LLMs in this study.
- Meditron-7b[17] is an adapted version of LLaMA-2 to the medical domain through continued pre-training on a comprehensively curated medical corpus, including selected PubMed papers and abstracts, a new dataset of internationally-recognized medical guidelines, and general domain data from RedPajama-v1[28]. It outperforms LLaMA-2-7B on multiple medical reasoning tasks.

[1] https://rarediseases.org/

Table 2: Zero-shot performance of diverse LLMs on token-level NER.

Model	Specificity	Source	Disease			Rare Disease		
			P	R	F1	P	R	F1
LLaMA-2	General	Open-Source	0.0564	0.0149	0.0219	0.3548	0.1066	0.1494
Meditron	Medical	Open-Source	0.0586	0.0242	0.0318	0.3409	0.1493	0.1877
UniversalNER	Task-Specific	Open-Source	0.1544	0.0731	0.0936	0.5493	0.1634	0.2267
Llama2-MedTuned	Medical	Open-Source	0.1391	**0.7201**	0.2332	0.3988	**0.8080**	0.5340
ChatGPT-3.5	General	Proprietary	0.1173	0.3842	0.1798	**0.5810**	0.2513	0.3509
ChatGPT-4	General	Proprietary	**0.1993**	0.4377	**0.2739**	0.5767	0.5761	**0.5764**

- UniversalNER-7b[14] is an instruction-fine-tuned version of LLaMA-7b[13] which is specified in named entity recognition. Essentially, the authors prompt ChatGPT to generate instruction-tuning data for NER from web text and conduct instruction-tuning on LLaMA. UniversalNER shows state-of-the-art document-level NER performance on multiple benchmarks across diverse domains, including biomedicine.
- Llama2-MedTuned-7b[18] is a medically-adapted version of the LLaMA-2-7b[16] model. Essentially, the authors specifically create a medical instruction-based dataset consisting of approximately $200,000$ samples covering clinical NLP tasks including token-level named entity recognition, relation extraction, document classification, question answering, natural language inference, and conduct instruction-tuning on LLaMA. Llama2-MedTuned demonstrates on-par performance with domain-specific models like BioClinicalBERT[8] on several benchmarks.
- ChatGPT-3.5 and ChatGPT-4 are arguably by far the most capable LLMs and have demonstrated superior performance and robustness across various domains. However, their proprietary nature limits their use in the clinical domain, where data privacy is of utmost importance. Specifically, we use `gpt-35-turbo-1106` and `gpt-4-1106-preview` in the experiments[2].

Learning Strategies In this study, we consider both in-context learning and instruction-fine-tuning to further adapt the LLMs to the token-level NER task. For in-context learning, we employ few-shot learning and retrieval-augmented generation (RAG), the prompt templates of which are shown in Figure 2. Essentially, we randomly select 5 samples from the training set as the demonstration to the LLMs. To ensure consistency, we fix the 5 samples for all few-shot experiments in the study. For RAG implementation, we utilize the LlamaIndex framework, incorporating NORD rare disease articles as the knowledge base. Specifically, these articles are segmented into chunks ($chunk\_size = 512$), embedded using the `bge-large-en-v1.5` model, and stored in a vector database. In the retrieval stage, we convert a query using the embedding model and retrieve the top-k ($similarity\_top\_k = 2$) most relevant text chunks based on similarity with the query embedding. The retrieved chunks are then integrated into the LLM prompts for enriched context.

For instruction-fine-tuning, we follow previous work[18] and convert the RareDis dataset to the Stanford Alpaca[29] format. This format is structured for instruction-following, comprising three parts: an instruction, an input, and an output, as shown in Figure 2. We then fine-tune the Llama2-MedTuned-7b model on the transformed RareDis dataset. We use LoRA adapters[30] for the fine-tuning. It is worth noting that while instruction-fine-tuning (or instruction-tuning) somewhat contradicts our initial intent due to its reliance on sufficient data, it remains crucial for our investigation into LLMs' performance capabilities and limitations in comparison to smaller BERT-like models in data-rich scenarios. Further implementation details are available in our code repository[3].

Evaluation We utilize Precision (P), Recall (R), and F1-score (F1) as our evaluation metrics for the token-level NER task. For models designed for document-level NER, such as UniversalNER-7b[14], we adapt their outputs to token-level by exact string matching. Initially, we assess a range of LLMs to investigate their overall performance. This is followed

[2]We access the OpenAI models through Microsoft Azure's HIPAA-certified platform.

[3]https://github.com/qiuhaolu/tner

Table 3: In-depth performance analysis of ChatGPT and Llama2-MedTuned across different entity types.

Prompting	Entity Type	ChatGPT-3.5			ChatGPT-4			Llama2-MedTuned		
		P	R	F1	P	R	F1	P	R	F1
zero-shot	Disease	0.1173	0.3842	0.1798	0.1993	0.4377	0.2739	0.1391	0.7201	0.2332
	Rare Disease	0.5810	0.2513	0.3509	0.5767	0.5761	0.5764	0.3988	0.8080	0.5340
	Skin Rare Disease	0.1200	0.0833	0.0984	0.4762	0.1389	0.2151	0.0686	0.8542	0.1270
	Sign	0.1314	0.0844	0.1028	0.1802	0.2665	0.2150	0.1712	0.3364	0.2269
	Symptom	0.0305	0.5882	0.0580	0.0574	0.7255	0.1063	0.0123	0.4902	0.0240
	Average	0.1960	0.2783	0.1580	0.2980	0.4289	0.2773	0.1580	0.6418	0.2290
few-shot	Disease	0.1871	0.2723	0.2218	0.2254	0.2977	0.2566	0.1517	0.5038	0.2332
	Rare Disease	0.5615	0.5512	0.5563	0.6011	0.5901	0.5955	0.5106	0.8069	0.6254
	Skin Rare Disease	0.1231	0.3958	0.1878	0.2953	0.3958	0.3383	0.0823	0.8333	0.1498
	Sign	0.1681	0.0813	0.1096	0.2749	0.2500	0.2619	0.1661	0.3283	0.2206
	Symptom	0.0478	0.6471	0.0889	0.0900	0.5490	0.1547	0.0140	0.5294	0.0273
	Average	0.2175	0.3895	0.2329	0.2973	0.4165	0.3214	0.1849	0.6003	0.2513
RAG	Disease	0.1568	0.2952	0.2048	0.2211	0.3995	0.2847	0.1410	0.6412	0.2312
	Rare Disease	0.6625	0.4595	0.5427	0.6552	0.5534	0.6000	0.4813	0.7918	0.5987
	Skin Rare Disease	0.1333	0.0833	0.1026	0.7419	0.1597	0.2629	0.0886	0.7986	0.1595
	Sign	0.1586	0.0607	0.0807	0.2263	0.2037	0.2144	0.0955	0.0802	0.0872
	Symptom	0.0420	0.6078	0.0786	0.0482	0.7451	0.0906	0.0173	0.4314	0.0314
	Average	0.2306	0.3013	0.2033	0.3785	0.4123	0.2905	0.1645	0.5486	0.2216

by a more focused evaluation of those LLMs that demonstrate proficiency in token-level NER. Subsequently, we conduct a detailed analysis of the aforementioned learning strategies, examining their influence on the performance of the LLMs. Additionally, we perform an error analysis to gain insights into the common errors across different models and learning strategies.

Results

LLM Initial Evaluation In our initial evaluation, we focus on assessing the performance of the aforementioned LLMs in token-level NER using the RareDis-v1 dataset, particularly targeting the two key target entities, i.e., `Disease` and `Rare Disease`. The results in Table 2 show that, generally, all the models struggle with this task, as evidenced by their modest performance metrics. Notably, Meditron, with its medical training, demonstrates marginally better results than LLaMA-2. However, even UniversalNER, which is optimized for document-level NER, does not show significantly improved performance in this token-level task. In contrast, Llama2-MedTuned-7b demonstrates promising results, closely trailing behind ChatGPT-4, which highlights the potential of local open-source LLMs in clinical NLP applications, specifically in token-level NER.

Focused Evaluation of Token-Level NER Capable LLMs Table 3 provides an in-depth performance analysis of ChatGPT-3.5, ChatGPT-4, and Llama2-MedTuned, selected based on their promising results in our initial evaluation. We encompass a range of experimental settings in this subsection, including zero-shot, few-shot, and retrieval-augmented generation (RAG). The discussion on the fine-tuning approach is reserved for the following subsection, as it requires the use of the entire training dataset.

A notable observation is the relatively high performance in identifying `Rare Disease`, likely due to its uniqueness and specificity. However, the models have varying degrees of difficulty in identifying subjective experiences reported by the patient, i.e., `Symptom`, as reflected by their close-to-zero precision and F1-scores. Moreover, we show that few-shot learning can effectively improve the performance across all models, with an increase ranging from 3% to 8% in average F1 scores. We also observe that ChatGPT-3.5 shows a substantial improvement of 20% in the F1-score for

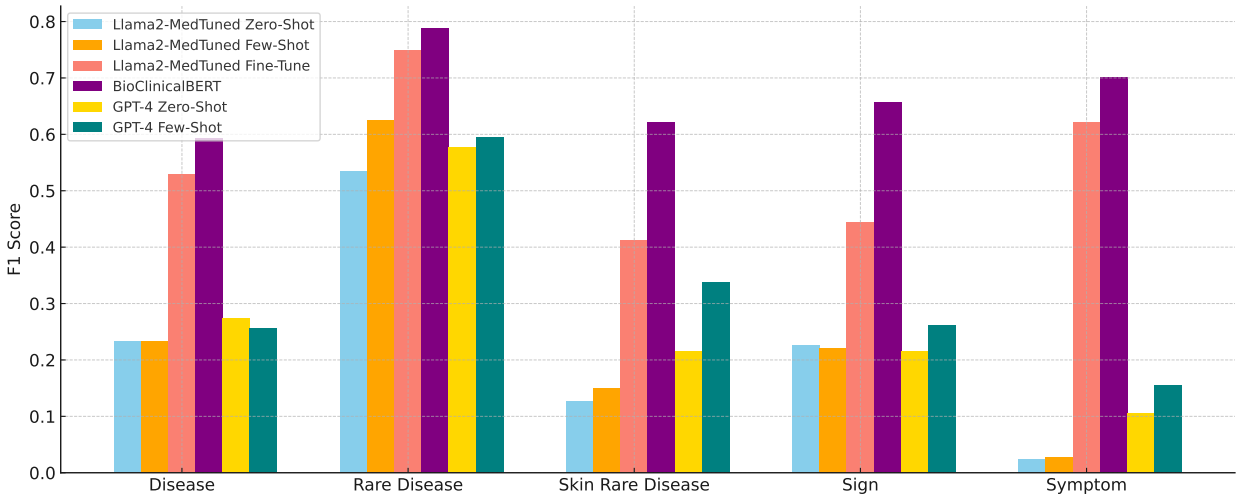


Figure 3: Performance evaluation of Llama2-MedTuned, BioClinicalBERT, and ChatGPT-4 across zero-shot, few-shot, and fine-tuning methods.

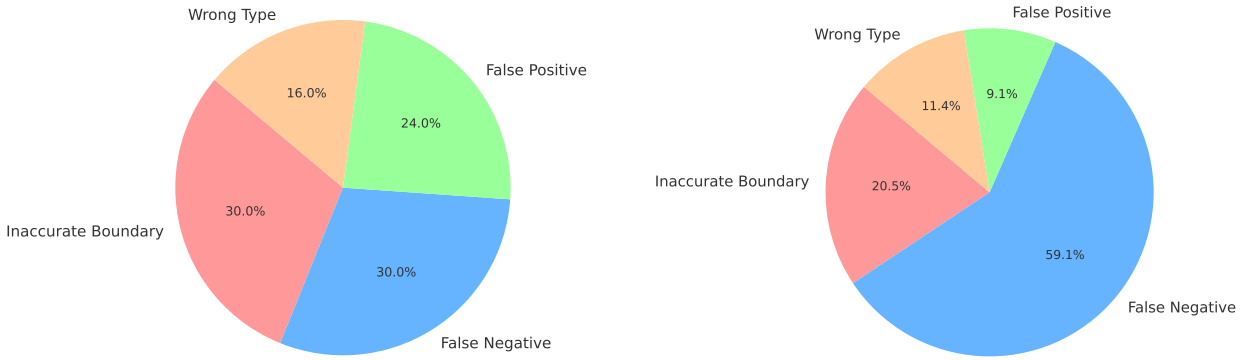
`Rare Disease`. These observations suggest that few-shot learning serves as an effective method to enhance LLMs' performance in token-level NER.

On the other hand, while RAG shows some promise, its overall impact on enhancing model performance in this task is limited. The data shows that RAG particularly benefits the identification of `Rare Disease` and `Skin Rare Disease` across all three models. This improvement could be attributed to RAG's ability to leverage external knowledge bases, i.e., NORD articles about rare diseases, which might be rich in rare disease terminology. However, for more general and common entities like `Disease`, `Symptom` and `Sign`, RAG's advantages are less impressive. This highlights the need for further refinement of RAG methods in clinical NLP applications.

Besides zero-shot prompting and in-context learning, we also investigate the impact of instruction-fine-tuning on these LLMs in this task. As shown in Table 2 and Table 3, despite the performance gain via few-shot learning, the overall performance of prompting LLMs is far from satisfactory. In this experiment, we aim to investigate LLMs' performance capabilities and limitations compared to fine-tuned BERT-like models under data-rich conditions, which is the common solution to NER in current practice. The results are shown in Figure 3. Essentially, Figure 3 visualizes the F1 scores of various models, including BioClinicalBERT fine-tune, Llama2-MedTuned under different learning strategies (zero-shot, few-shot, fine-tune), and ChatGPT-4 in zero-shot and few-shot scenarios. Specifically, the performance of Llama2-MedTuned models demonstrates significant effectiveness, not only outperforming ChatGPT-4 in most scenarios but also closely rivaling BioClinicalBERT when fine-tuned. This highlights the potential of open-source LLMs in this task.

Error Analysis We present an error analysis to further understand the mistakes made by the LLMs in this task. Essentially, we randomly select 50 sentences from the test dataset and manually categorize the errors for each incorrect prediction. We investigate the two best-performing models, i.e., ChatGPT-4 under few-shot learning and Llama2-MedTuned under fine-tuning. We identify 4 types of errors in the experiment, i.e., Inaccurate Boundary, where the model incorrectly identifies the start or end of an entity; Wrong Type, where the entity is recognized, but its type is misclassified; False Negative, where the model overlooks an entity present in the text; and False Positive, where the model mistakenly identifies a non-entity as an entity.

The results in Figure 4 show that Inaccurate Boundary errors are more prevalent for ChatGPT-4, suggesting challenges in precisely detecting entity boundaries. In contrast, Llama2-MedTuned demonstrates a much higher chance of False Negatives, indicating its inability to identify entities that are present. Both models show fewer Wrong Type and False Positive errors, reflecting a relatively stronger performance in correctly classifying entities and avoiding over-detection. Insights from these error patterns highlight the strengths and weaknesses of each model in token-level NER and indicate directions for future improvements, especially in fine-tuning local open-source LLMs.



(a) ChatGPT-4 **(b)** Llama2-MedTuned

Figure 4: Comparative error analysis of ChatGPT-4 and Llama2-MedTuned.

Discussion and Conclusion

In this study, we investigate the capabilities of LLMs in token-level NER for rare diseases and their associated phenotypes on the RareDis-v1 dataset. Essentially, we find that in general, most LLMs struggle with this challenging task. We also find that a medically-adapted LLaMA-2-7b model, i.e., Llama2-MedTuned, surpasses the performance of ChatGPT-3.5 and matches the performance of ChatGPT-4 on this task, highlighting the potential of local open-source LLMs in clinical NLP applications. Our further analysis suggests that via in-context learning, i.e., few-shot learning and RAG, the performance can be improved while RAG has a relatively limited influence. We also show that after fine-tuning Llama2-MedTuned on this dataset, its performance gets boosted and significantly outperforms that of ChatGPT-4, approaching the levels of BioClinicalBERT. It is worth noting that we use LoRA adapters to perform the fine-tuning, instead of the entire model. Based on these results, we anticipate that full model fine-tuning could potentially enable the Llama2-MedTuned to match BioClinicalBERT's performance.

Large language models (LLMs) have demonstrated superior performance in question-answering-related tasks across almost all domains, including the clinical field. However, their performance in conventional NLP tasks, such as information extraction, has been questioned. Unlike existing studies that either focus on document-level NER or solely rely on the prompt engineering of ChatGPT-series models, our experiment covers a broad scope of LLMs, from proprietary to local open-source and from general-purpose to medically adapted models. We thoroughly evaluate the LLMs on the token-level NER task, showing their struggle and potential directions for refinement. This study could shed light on adapting LLMs to specific NLP applications in clinical settings.

There are several factors that contribute to the struggle of LLMs on this task. Firstly, token-level NER is more challenging than document-level NER. Token-level NER is a token classification task and places significant emphasis on the representation of each individual token. This is in contrast to document-level NER which is a sequence classification task where a pooled representation is often sufficient for predictions, making the role of individual token representation less critical. Secondly, the foundational pre-training objective of most LLMs (decoder-only transformers) is causal language modeling, i.e., next token prediction. Unlike encoder-only transformers like BERT, LLMs don't have an encoder structure that is pre-trained to maximize the model's text representation power, making them less effective in classification tasks. Thirdly, token-level NER often involves a variety of special, previously unseen symbols, such as the diverse BIO tags. These unique elements can present additional complexities for LLMs to understand and operate.

There are also a few options to adapt these LLMs to the token-level NER task. A typical solution is continuous pre-training, or instruction-fine-tuning, just as Llama2-MedTuned and what we do using the rare disease data. Basically, the idea is to cast the NER problem as text generation, so the LLM can have a better understanding of the task and thus process it correctly in the same manner as it was pre-trained. For instance, Yang *et al.* integrate Human Phenotype Ontology (HPO) labels into clinical notes and fine-tune GPT-based models (e.g., GPT-J) for phenotype recognition[31].

They also conduct a comprehensive comparative analysis of encoder-based and decoder-based models for this task. Another solution is to leverage LLMs to generate synthetic data and use the data to train smaller specified models. In addition to the data, one can also alter the model architecture of LLMs. For example, Li *et al.* propose to fine-tune LLaMA-2 models in a label-supervised way, by removing the causal mask from the decoders to enable bidirectional self-attention which is essential for token classification tasks like NER[32]. However, in our preliminary experiment, the method does not work as expected and the performance is worse than instruction-fine-tuning on the rare disease dataset. We leave it for future study.

This study has certain limitations. Firstly, it does not explore the potential benefits of specific prompt engineering techniques, such as utilizing feedback from error analysis to refine prompts and thereby enhance model performance. Additionally, our approach mainly focuses on prompting LLMs to generate BIO tags. However, there are alternative approaches that are unexplored, such as the generation of text marked by special symbols to represent span information, which could offer a different perspective on model efficacy. Furthermore, in terms of instruction-fine-tuning of LLMs, our methodology is limited to employing LoRA adapters. A more extensive approach such as full model fine-tuning remains untested and could be valuable for future research.

To conclude, in this study, we explore the capabilities and limitations of existing LLMs, especially local open-source LLMs, in the task of token-level NER of rare diseases and their associated phenotypes. Through this exploration, we aim to contribute to the advancement of LLM-based token-level NER methodologies in the clinical domain, ultimately aiding in the improvement of patient care by enhancing the processing and understanding of clinical texts.

Acknowledgments This work was conducted under support from the National Human Genome Research Institute (R01HG12748).

References

1. Henry J, Pylypchuk Y, Searcy T, Patel V, et al. Adoption of electronic health record systems among US non-federal acute care hospitals: 2008–2015. ONC data brief. 2016;35(35):2008-15.
2. Kundeti SR, Vijayananda J, Mujjiga S, Kalyan M. Clinical named entity recognition: Challenges and opportunities. In: 2016 IEEE International Conference on Big Data (Big Data). IEEE; 2016. p. 1937-45.
3. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association. 2010;17(3):229-36.
4. Denny JC, Irani PR, Wehbe FH, Smithers JD, Spickard III A. The KnowledgeMap project: development of a concept-based medical school curriculum database. In: AMIA Annual Symposium Proceedings. vol. 2003. American Medical Informatics Association; 2003. p. 195.
5. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association. 2010;17(5):507-13.
6. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS); 2017. p. 6000-10.
7. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (NAACL-HLT); 2019. p. 4171-86.
8. Alsentzer E, Murphy J, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly Available Clinical BERT Embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. Minneapolis, Minnesota, USA: Association for Computational Linguistics; 2019. p. 72-8. Available from: https://www.aclweb.org/anthology/W19-1909.
9. Lewis P, Ott M, Du J, Stoyanov V. Pretrained Language Models for Biomedical and Clinical Tasks: Understanding and Extending the State-of-the-Art. In: Proceedings of the 3rd Clinical Natural Language Processing Workshop. Online: Association for Computational Linguistics; 2020. p. 146-57. Available from: https://www.aclweb.org/anthology/2020.clinicalnlp-1.17.
10. Sung M, Jeong M, Choi Y, Kim D, Lee J, Kang J. BERN2: an advanced neural biomedical named entity recognition and normalization tool. Bioinformatics. 2022;38(20):4837-9.

11. Dharssi S, Wong-Rieger D, Harold M, Terry S. Review of 11 national policies for rare diseases in the context of key patient needs. Orphanet journal of rare diseases. 2017;12(1):1-13.
12. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nature medicine. 2023;29(8):1930-40.
13. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:230213971. 2023.
14. Zhou W, Zhang S, Gu Y, Chen M, Poon H. UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition. In: The Twelfth International Conference on Learning Representations; 2023. .
15. Shyr C, Hu Y, Bastarache L, Cheng A, Hamid R, Harris P, et al. Identifying and Extracting Rare Diseases and Their Phenotypes with Large Language Models. Journal of Healthcare Informatics Research. 2024:1-24.
16. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:230709288. 2023.
17. 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 arXiv:231116079. 2023.
18. Rohanian O, Nouriborji M, Clifton DA. Exploring the Effectiveness of Instruction Tuning in Biomedical Language Processing. arXiv preprint arXiv:240100579. 2023.
19. Chung HW, Hou L, Longpre S, Zoph B, Tay Y, Fedus W, et al. Scaling instruction-finetuned language models. arXiv preprint arXiv:221011416. 2022.
20. Hu Y, Chen Q, Du J, Peng X, Keloth VK, Zuo X, et al. Improving large language models for clinical named entity recognition via prompt engineering. Journal of the American Medical Informatics Association. 2024:ocad259.
21. Li Y, Li J, He J, Tao C. AE-GPT: using large language models to extract adverse events from surveillance reports-a use case with influenza vaccine adverse events. Plos one. 2024;19(3):e0300919.
22. Li Y, Viswaroopan D, He W, Li J, Zuo X, Xu H, et al. Improving Entity Recognition Using Ensembles of Deep Learning and Fine-tuned Large Language Models: A Case Study on Adverse Event Extraction from Multiple Sources. arXiv preprint arXiv:240618049. 2024.
23. Lu Q, Dou D, Nguyen TH. Parameter-Efficient Domain Knowledge Integration from Multiple Sources for Biomedical Pre-trained Language Models. In: Moens MF, Huang X, Specia L, Yih SWt, editors. Findings of the Association for Computational Linguistics: EMNLP 2021. Punta Cana, Dominican Republic: Association for Computational Linguistics; 2021. p. 3855-65. Available from: https://aclanthology.org/2021.findings-emnlp.325.
24. Lu Q, Dou D, Nguyen T. ClinicalT5: A Generative Language Model for Clinical Text. In: Goldberg Y, Kozareva Z, Zhang Y, editors. Findings of the Association for Computational Linguistics: EMNLP 2022. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics; 2022. p. 5436-43. Available from: https://aclanthology.org/2022.findings-emnlp.398.
25. Nguyen CQ, Kariyawasam D, Alba-Concepcion K, Grattan S, Hetherington K, Wakefield CE, et al. 'Advocacy groups are the connectors': Experiences and contributions of rare disease patient organization leaders in advanced neurotherapeutics. Health Expectations. 2022;25(6):3175-91.
26. Martínez-deMiguel C, Segura-Bedmar I, Chacón-Solano E, Guerrero-Aspizua S. The RareDis corpus: a corpus annotated with rare diseases, their signs and symptoms. Journal of Biomedical Informatics. 2022;125:103961.
27. Segura-Bedmar I, Camino-Perdones D, Guerrero-Aspizua S. Exploring deep learning methods for recognizing rare diseases and their clinical manifestations from texts. BMC bioinformatics. 2022;23(1):263.
28. Computer T. RedPajama: an Open Dataset for Training Large Language Models; 2023. Available from: https://github.com/togethercomputer/RedPajama-Data.
29. Taori R, Gulrajani I, Zhang T, Dubois Y, Li X, Guestrin C, et al.. Stanford Alpaca: An Instruction-following LLaMA model. GitHub; 2023. https://github.com/tatsu-lab/stanford_alpaca.
30. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: Low-Rank Adaptation of Large Language Models. In: International Conference on Learning Representations; 2022. Available from: https://openreview.net/forum?id=nZeVKeeFYf9.
31. Yang J, Liu C, Deng W, Wu D, Weng C, Zhou Y, et al. Enhancing phenotype recognition in clinical notes using large language models: PhenoBCBERT and PhenoGPT. Patterns. 2024;5(1).
32. Li Z, Li X, Liu Y, Xie H, Li J, Wang Fl, et al. Label supervised llama finetuning. arXiv preprint arXiv:231001208. 2023.