Envisioning the Future of Health Informatics and Digital Health
J. Mantas et al. (Eds.)

doi:10.3233/SHTI250051

Evaluation of the Performance of a Large Language Model to Extract Signs and Symptoms from Clinical Notes

C. Mahony REATEGUI-RIVERA[a,1] and Joseph FINKELSTEIN[a]
[a] *Department of Biomedical Informatics, School of Medicine, University of Utah, Salt Lake City, Utah.*
ORCiD ID: C. Mahony Reategui-Rivera https://orcid.org/0000-0002-4030-8777

Abstract. Large language models (LLMs) have increasingly been used to extract critical information from unstructured clinical notes, which often include important details not captured in the structured sections of electronic health records (EHRs). This study assesses the performance of the GPT-4o LLM in extracting signs and symptoms (S&S) from clinical notes, focusing on both general and organ-specific (urological and cardiorespiratory) contexts. Clinical notes from the MTSamples corpora were manually annotated for comparison with the S&S extraction results using LLM. GPT-4o was applied to extract S&S using named entity recognition techniques. Key performance metrics—precision, recall, and F1-score—were used to evaluate and compare general and organ-specific results. The model showed high precision in general S&S extraction (78%) and achieved the highest precision for organ-specific tasks in the cardiorespiratory dataset (87%). For the urinary dataset, precision was also strong (81%), with balanced recall and F1-scores across analyses. These findings underscore GPT-4o's effectiveness in both general and domain-specific S&S extraction but highlight the need for domain-specific tuning and optimization to further improve recall and generalizability in specialized medical contexts.

Keywords. Large Language Models, Natural Language Processing, Named Entity Recognition, Signs and Symptoms

1. Introduction

Recent advancements in natural language processing (NLP) have enabled the development of models capable of interpreting the free-text components of clinical notes. These notes, often composed of unstructured data, pose challenges due to variability and lack of standardization, yet they contain crucial details about a patient's symptoms, diagnosis, and treatment that may not be captured in the structured sections of electronic health records (EHRs) [1,2]. Unstructured clinical data, particularly in clinical notes, is especially valuable for managing conditions like cancer, COVID-19, cognitive disorders, and other health issues [2–4].

Large language models (LLMs) have significantly improved the ability to analyze such unstructured data, especially for extracting essential medical information from EHRs. Models like BERT and its biomedical variants (e.g., EHR-BERT, BioBERT,

[1] Corresponding Author: C. Mahony Reategui-Rivera, E-mail: mahony.reategui@utah.edu.

ClinicalBERT, and Symptom-BERT) have shown great success in extracting signs and symptoms (S&S) from clinical notes, with Symptom-BERT achieving high accuracy in symptom detection for chronic diseases [3,5]. However, these models often rely on task-specific annotations and training, which limits their generalizability [3].

In contrast, more advanced models like GPT-4o, with larger parameter sets and broader context windows, offer the potential for superior performance in extracting S&S from clinical notes without extensive task-specific training [6]. For instance, GPT-3.5 achieved 89% accuracy in extracting pathological classifications, surpassing traditional NLP approaches [6]. Given the enhanced capabilities of models like GPT-4o, their performance in clinical NLP tasks warrants further exploration.

Leveraging LLMs to detect S&S from clinical notes could revolutionize public health systems by providing critical information for clinical decision-making and improving predictive artificial intelligence models [7]. Therefore, this study aims to evaluate the ability of a modern LLM to extract S&S from a corpus of de-identified clinical notes.

2. Methods

2.1. Data Sources

The study utilized clinical notes from the MTSamples corpus (Medical Transcription Samples at www.mtsamples.com) focusing on notes related to urological and cardiorespiratory conditions. After an initial keyword-based filtering process, notes mentioning relevant urinary or cardiorespiratory pathologies were manually reviewed. Notes lacking information on S&S were excluded. The final dataset comprised 97 cardiorespiratory notes and 27 urological notes, with no further text preprocessing applied.

2.2. Expert Annotation

Manual clinical expert annotations were used as the gold standard for comparison reference with the LLM results. Annotations were standardized to ensure consistency, consolidating synonyms (e.g., "hematuria" and "blood in urine") under a single representative term. The standardization followed the physician's judgment, without formal validation against medical ontologies.

2.3. LLM Annotation

GPT-4o was used for S&S extraction via named entity recognition. The Chatbot Arena platform facilitated model interaction, and each clinical note was independently input as a prompt. Three prompting techniques were evaluated: (1) basic prompt with term definitions, (2) Auto-Chain of Thought, and (3) Chain of Thought with one-shot inference. The best-performing technique (Chain of Thought with one-shot inference) was applied throughout the analysis. Model memory was cleared between runs, and the extracted S&S were compared to the human annotations. Additional S&S identified by GPT-4o were considered but not further validated.

2.4. Analysis Plan

Model performance was assessed using precision, recall, and F1-score, with human annotations as the gold standard. A focused analysis was conducted for organ-specific S&S, with datasets filtered to include only urinary or cardiorespiratory-related S&S. Discrepancies between human and GPT-4o annotations were resolved by assuming the human annotations were correct, without further expert review.

3. Results

3.1. Urinary Dataset

For the urinary dataset, the model's general precision was 0.59, with a recall of 0.75 and an F1-score of 0.66. For urinary-specific S&S, the GPT-4o model showed substantial improvements, with precision rising to 0.81 and F1-score increasing to 0.76. However, recall slightly decreased to 0.72 compared to 0.75 previously. This suggests a significant reduction in false positives and a well-maintained recall rate, highlighting the model's ability to accurately identify relevant specific S&S in this domain.

3.2. Cardiorespiratory Dataset

For general S&S, the precision was 0.78, recall of 0.71, and a resulting F1-score of 0.74, reflecting balanced performance. For cardiorespiratory-specific S&S, precision rose significantly to 0.87, and the F1-score increased to 0.72, despite a decrease in recall of 0.62. Despite a lower recall, the improved precision suggests accurate identification of relevant S&S while minimizing false positives.

3.3. Comparison between Organ-specific Datasets

When comparing organ-specific performance, the model showed higher precision for cardiorespiratory S&S (0.87) compared to urinary S&S (0.81). Conversely, recall was higher for the urinary S&S compared to cardiorespiratory-specific S&S. The F1-scores were 0.72 for cardiorespiratory and 0.76 for urinary S&S, reflecting better overall performance in the urinary dataset.

Table 1. Performance Metrics for General and Organ-Specific (Cardiorespiratory and Urinary) Signs & Symptoms Extraction using GPT-4o.

	Urinary Dataset		**Cardiorespiratory Dataset**	
	General S&S	**Organ-specific S&S**	**General S&S**	**Organ-specific S&S**
Total TP	103	38	443	171
Total FP	72	9	127	26
Total FN	35	15	179	105
Precision	0.59	0.81	0.78	0.87
Recall	0.75	0.72	0.71	0.62
F1-Score	0.66	0.76	0.74	0.72

4. Discussion

4.1. Summary of Main Findings

The GPT-4o model exhibited a strong performance in extracting S&S from clinical notes, with differences across clinical specialties. For the urinary dataset, general S&S extraction showed balanced precision and recall, while urinary-specific S&S extraction achieved high precision and F1-score, reflecting a substantial reduction in false positives. In the cardiorespiratory dataset, general S&S extraction performed consistently (F1-score of 0.74), and cardiorespiratory-specific S&S extraction achieved the highest precision (0.87) but with lower recall (0.62). These results highlight GPT-4o's effectiveness in domain-specific tasks while maintaining strong general performance.

4.2. Comparison with Other Studies

The findings align with previous studies using LLMs for symptom extraction, which report high precision and F1-score but challenges with recall [1,8]. LLMs generally perform better in broader contexts and often require fine-tuning to handle specialized fields effectively [1]. Similar challenges observed in organ-specific S&S extraction have been noted in earlier studies, where recall for nuanced symptoms remained a limitation [8]. Domain-specific tuning, as seen in prior work, may be a solution to improve performance in areas such as urinary and cardiorespiratory symptoms.

Previous studies demonstrated comparable accuracy for S&S extraction using rule-based and machine learning approaches [9]. However, previous approaches required labor-intensive and time-consuming efforts for generation of large representative training datasets [10]. In addition, to achieve sufficient accuracy, these algorithms had frequently to be retrained for institutions to account for local contexts [11]. The LLM approach makes these steps unnecessary which greatly scales up the broad implementation of NLP pipelines.

4.3. Possible Explanation of Findings

The high precision observed for organ-specific S&S, particularly in the cardiorespiratory dataset, likely reflects the model's ability to recognize well-defined domain-specific patterns. However, the lower recall suggests that some nuanced or less frequent terms may still be missed, particularly in the cardiorespiratory dataset. The urinary dataset's more balanced performance may be attributed to the narrower scope of organ-specific S&S, which allowed the model to generalize better. For general S&S extraction, the consistent balance between precision and recall indicates the model's robustness in identifying broader clinical terminology.

4.4. Strengths and Limitations

This study has several strengths, including its dual focus on general and organ-specific S&S extraction, offering a comprehensive view of GPT-4o's performance across different contexts. The use of human-labeled data as the gold standard enhances the clinical relevance of the findings, and analyzing both precision and recall across domains provides valuable insight into the model's capabilities. However, the study has

limitations. The lower recall in organ-specific S&S, particularly in the cardiorespiratory dataset, underscores challenges in extracting specialized S&S without domain-specific tuning. The single annotator approach may have introduced bias, though standardized rules were used to mitigate this; future studies may consider double annotation. Additionally, the smaller urinary dataset may have GPT-4o's generalizability. The exclusion of procedural and surgical notes, while necessary for focusing on relevant S&S, may have limited the scope of the extracted data.

4.5. Conclusion

This study demonstrates that the GPT-4o model is highly effective in extracting general and organ-specific S&S, excelling in precision for domain-specific tasks. Urinary-specific S&S extraction achieved the most balanced performance, while cardiorespiratory-specific S&S extraction showed the highest precision. These results highlight GPT-4o's potential for clinical NLP tasks, though domain-specific tuning and further optimization are necessary to improve recall and generalizability, particularly in organ-specific contexts.

References

[1] Vithanage D, Zhu Y, Zhang Z, et al. Extracting Symptoms of Agitation in Dementia from Free-Text Nursing Notes Using Advanced Natural Language Processing. Stud Health Technol Inform. 2024;310:700–704.
[2] Van Olmen J, Van Nooten J, Philips H, et al. Predicting COVID-19 Symptoms From Free Text in Medical Records Using Artificial Intelligence: Feasibility Study. JMIR Med Inform. 2022;10(4):e37771.
[3] Zeinali N, Albashayreh A, Fan W, et al. Symptom-BERT: Enhancing Cancer Symptom Detection in EHR Clinical Notes. J Pain Symptom Manage. 2024;68(2):190-198.e1.
[4] Jethani N, Jones S, Genes N, et al. Evaluating ChatGPT in Information Extraction: A Case Study of Extracting Cognitive Exam Dates and Scores [Internet]. medRxiv; 2023 [cited 2024 Sep 16]. p. 2023.07.10.23292373. Available from: https://www.medrxiv.org/content/10.1101/2023.07.10.23292373v1.
[5] Niu H, Omitaomu OA, Langston MA, et al. EHR-BERT: A BERT-based model for effective anomaly detection in electronic health records. J Biomed Inform. 2024;150:104605.
[6] Huang J, Yang DM, Rong R, et al. A critical assessment of using ChatGPT for extracting structured data from clinical notes. npj Digit Med. 2024;7(1):1–13.
[7] Lee S, Kim H-S. Prospect of Artificial Intelligence Based on Electronic Medical Record. J Lipid Atheroscler. 2021;10(3):282–290.
[8] Taub-Tabib H, Shamay Y, Shlain M, et al. Identifying symptom etiologies using syntactic patterns and large language models. Sci Rep. 2024;14(1):16190.
[9] Shah-Mohammadi F, Cui W, Finkelstein J. Entity Extraction for Clinical Notes, a Comparison Between MetaMap and Amazon Comprehend Medical. Stud Health Technol Inform. 2021;281:258–262.
[10] Shah-Mohammadi F, Cui W, Finkelstein J. Comparison of ACM and CLAMP for Entity Extraction in Clinical Notes. Annu Int Conf IEEE Eng Med Biol Soc. 2021;2021:1989–1992.
[11] Cui W, Shah-Mohammadi F, Finkelstein J. Using Electronic Medical Records and Clinical Notes to Predict the Outcome of Opioid Treatment Program. Stud Health Technol Inform. 2023;305:568–571.