Impact of Context on Large Language Models for Clinical Named Entity Recognition

Weipeng Zhou, PhD[1], Rui Shi[1], Gui Yang[1], Anran Li, PhD[1], Xuguang Ai, MS[1], Qingyu Chen, PhD[1], Yan Hu, PhD, Hua Xu, PhD[1], Timothy A. Miller, PhD[2]

[1]Yale University, New Haven, CT; [2]Boston Children's Hospital, Boston, MA; HT Health Houston

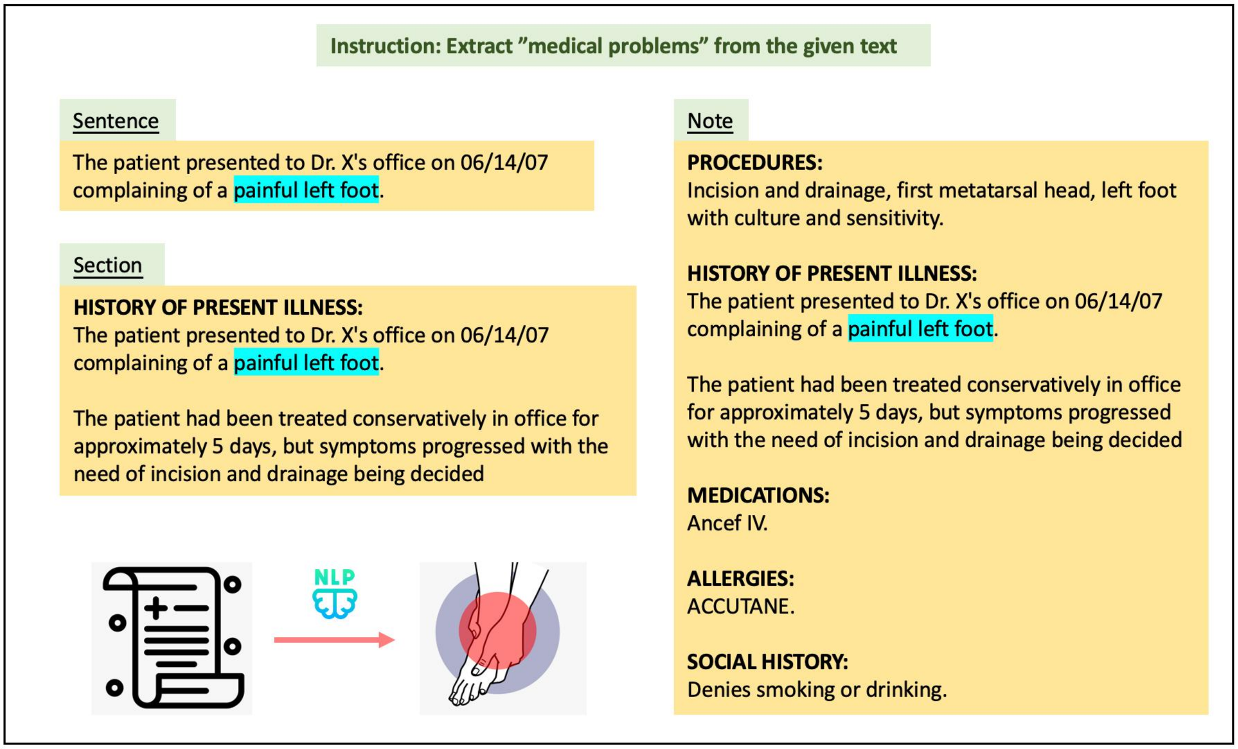
Abstract

This study explores the impact of context on Large Language Model (LLM) performance in clinical named entity recognition (NER), focusing on three context levels—sentence, section, and document—and two methods of context incorporation: (1) embedded context - expanding the input to include broader context (e.g., replacing a sentence with its full section or entire document) and (2) detached context - incorporating context as independent material before the input. Using the MTSamples and i2b2 datasets, we evaluated GPT-4o models across varying context conditions. Our analysis reveals that detached context at the section level enhanced model performance, achieving 0.594 exact match F1 on MTSamples and 0.485 on i2b2, while embedded context generally reduced performance.

Introduction

Clinical named entity recognition (NER) - the automated identification of medical entities like diagnoses, treatments, and tests in unstructured text - forms a critical foundation for healthcare natural language processing (NLP) systems[1]. Traditional approaches using specialized models like BioClinicalBERT[2] achieve strong performance but require extensive annotated data and face inherent limitations: constrained context windows (typically 512 tokens) and sentence-level processing that disrupts clinical narrative flow [2], [3]. The emergence of large language models (LLMs) presents new opportunities through their exceptional zero-shot capabilities and expanded context processing - modern systems like GPT-4o can analyze documents exceeding 100,000 tokens while maintaining clinical concept understanding [4], [5], [6].

While context utilization has been studied in conventional NER systems [7], [8], [9], [10], LLMs present unique challenges and opportunities. Early supervised approaches demonstrated that structural elements of clinical notes - particularly UMLS (Unified Medical Language System) semantic types within section headers - significantly improve disambiguation of ambiguous medical abbreviations (e.g., distinguishing "PE" as pulmonary embolism vs. physical exam) [10], but LLMs' capacity to leverage document-level context remains underexplored. This is an important question because, in recent years, the size of context that LLMs can process at once has greatly expanded, from 2048 tokens in early models [11] to 10 million tokens in some modern models [12]. This expansion in capability may tempt users to process entire documents (or datasets) in one prompt, and the consequences of such actions still lack understanding. Furthermore, clinical contexts can be presented through distinct organizational levels (sentence, section, note) and integration methods - either embedded within the target text or detached as prefatory information - each configuration potentially influencing extraction accuracy through different cognitive pathways.



Instruction: Extract "medical problems" from the given text

Sentence

The patient presented to Dr. X's office on 06/14/07 complaining of a painful left foot.

Section

HISTORY OF PRESENT ILLNESS:
The patient presented to Dr. X's office on 06/14/07 complaining of a painful left foot.

The patient had been treated conservatively in office for approximately 5 days, but symptoms progressed with the need of incision and drainage being decided

Note

PROCEDURES:
Incision and drainage, first metatarsal head, left foot with culture and sensitivity.

HISTORY OF PRESENT ILLNESS:
The patient presented to Dr. X's office on 06/14/07 complaining of a painful left foot.

The patient had been treated conservatively in office for approximately 5 days, but symptoms progressed with the need of incision and drainage being decided

MEDICATIONS:
Ancef IV.

ALLERGIES:
ACCUTANE.

SOCIAL HISTORY:
Denies smoking or drinking.

Figure 1. The diagram illustrates the type of context when adding context information for named entity recognition extraction using large language models. At the sentence level, the model's input contains a single sentence and no context is provided. At the section level, the input is expanded to include the entire section, providing section-specific context. At the note level, the input is expanded to include the entire note, providing note-level context. This method of providing context by expanding the original sentence is referred to as **embedded context** in this study.

Our study systematically evaluates context utilization strategies for clinical NER through three complementary lenses:

1. **Granularity:** We investigated the granularity of context at three distinct levels—sentence, section (e.g., History of Present Illness), and document-level context (entire clinical notes)—to determine the optimal scope of contextual information for accurate entity recognition.
2. **Integration**: We investigated two approaches for providing the context in the prompt: (a) embedded context (Figure 1), where the target input was expanded to include broader surrounding text (e.g., replacing a single sentence with its complete section or entire document), and (b) detached context, where context material was incorporated separately in the prompt before the target sentence, maintaining the original input's boundaries.
3. **Scale**: Quantitative assessment of the relationship between context volume (adding a varying number of sentences around input to provide ad-hoc context) and model performance, examining how increasing context size affects model performance.

Methods

Datasets

We evaluated our approach using two datasets with distinct characteristics: MTSamples and i2b2. The MTSamples [5] publicly available corpus contains synthetic notes annotated following the annotation guidelines from 2010 i2b2 challenge[1] on concepts, assertions, and relations in clinical text. It has entity-level labels (medical problems, treatments, tests) applied at sentence level. During annotation, clinical experts first extracted "History of Present Illness" sections from full notes and split into sentences, resulting in sentence-level context and section-level context available for later experiments. The MTSamples dataset contains 163 sections (History of Present Illness only) and 1004 sentences.

The i2b2 dataset we used in this study is a subset of the 2010 original i2b2 corpus[1]. The original i2b2 corpus contains discharge summaries from Beth Israel Deaconess Medical Center, annotated at full-note level. Since the original i2b2 corpus does not contain section type and boundary annotations, we leveraged section annotations from Tepper et al. [13] which annotated a subset of i2b2 notes. The section annotations covered 38 notes in i2b2 and we only used this subset for constructing the i2b2 dataset in this study. We further used SpaCy [14] to segment sections into sentences and words (tokens). The resulted i2b2 dataset contains 38 notes, 598 sections and 2838 sentences.

Table 1 shows summary statistics for the MTSamples and i2b2 datasets. At the sentence level, MTSamples and i2b2 have similar token counts. At the section level, the average token and sentence counts for MTSamples is substantially higher since MTSamples only contains the longer "History of Present Illness" sections. At the note level, i2b2 on average contains 16 sections, with each note having on average 1013 tokens.

Table 1. Token, section and sentence count for MTSamples and i2b2 aggregated at sentence, section and note level.

	Sentence	Section		Note	
Dataset	# Token	# Token	# Sentence	# Token	# Section
MTSamples	16±10	170±105	11±10	N/A	N/A
i2b2	13±12	64±126	5±8	1013±646	16±5

When evaluating large language models on MTSamples, we performed our experiments on the test set as provided by Hu et al. (5) where the dataset was split into 6:2:2 for training, validation and test set [5]. For the i2b2 dataset, we applied large language models on the full dataset due to the reduced corpus size after matching to the section annotations in Tepper et al.[13].

Model

We used GPT-4o, specifically the 2024-08-01-preview version with default hyperparameters[6]. GPT-4o is a cutting-edge large language model from OpenAI. It features a 128k token context window, enabling it to handle long-form text generation and complex reasoning tasks more effectively. The GPT-4o API was accessed through Yale University's Azure HIPAA-compliant environment, ensuring secure and privacy-preserving processing of sensitive data.

Experiments

Context evaluation

Baseline

We established baseline performance by evaluating GPT-4o (2024-08-01-preview with default hyperparameters), which has a context window of 128k, on the MTSamples and i2b2 datasets through sentence-level processing. Each input sentence was accompanied by a standardized prompt template (Appendix Section 1) and processed through the GPT-4o API. We reused the best performing named entity recognition prompt from Hu et al., which was obtained via extensive prompt engineering[5].The GPT-4o API was provided by Yale University Azure HIPAA-compliant environment. To ensure reliability, we conducted five independent trials under identical conditions, maintaining this replication protocol across subsequent experiments unless otherwise stated.

For comparative analysis, we benchmarked GPT-4o's zero-shot performance against supervised baseline models. For MTSamples dataset, the BioClinicalBERT [2] fine-tuning results were obtained from Hu et al. [5]. For i2b2 dataset, we randomly split the dataset into 7:1:2 for training, validation and test sets, and fine-tuned BioClinicalBERT [2] by setting the epoch size to 10, the learning rate to $5*10^{-5}$ with 5 and the batch size to 4.

Embedded context

In traditional supervised learning paradigm like BERT, a common approach to incorporating context information is by providing longer examples during training. Traditionally, BERT-like models suffer from input length limitations—most BERT models support input lengths of up to 512 tokens [2]. In contrast, large language models typically support significantly longer inputs. We investigated embedded-section and embedded-note context by expanding the sentence to a section and the full clinical note as input. The prompt template used in the embedded context study is the same as the one in the baseline.

Detached context

We developed a novel context integration approach where contextual information appears as instructional prefixes rather than integrated input components. This detached context methodology presents two key advantages over embedded approaches: (1) maintains baseline input length consistency (sentence-level processing), (2) preserves output sequence length, reducing the risk of erroneous outputs. For the prompt template, we made a slight change to the baseline prompt by inserting "### Context\nHere is the context for the following input text but do not extract entities from it" before the input appears (Appendix 2)

Model evaluation and metrics

To ensure the robustness of the experiments, prompt engineering and preliminary experiments were conducted exclusively on the training splits of the MTSamples dataset. For the MTSamples dataset, results were reported on the held-out test set defined by Hu et al. [5]. For the i2b2 dataset, prompt development was not performed due to its limited size; instead, we reused the prompt developed for MTSamples and evaluated model performance on the i2b2 test set. Each evaluation was repeated five times, with results reported as the mean and standard deviation of the metrics.

We evaluated our models using precision, recall, and F1 scores under strict- and relaxed-match criteria. Precision is calculated as the number of true positives (TP) divided by the sum of true positives and false positives (FP). Recall is calculated as the number of true positives (TP) divided by the sum of true positives and false negatives (FN). The F1 score is the harmonic mean of precision and recall. A true positive (TP) denotes a case correctly predicted as positive, while a false positive (FP) represents a case incorrectly predicted as positive. A true negative (TN) refers to a case correctly predicted as negative, and a false negative (FN) indicates a case incorrectly predicted as negative.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

In the strict-match setting, predicted entities were considered correct only if their spans exactly matched the gold-standard boundaries. In the relaxed-match setting, predicted entities were considered correct if there was any partial overlap between the predicted and annotated spans.

Context analysis

Model performance with varying context sizes

To further investigate how contextual information quantity affects model performance, we implemented a progressive window expansion paradigm. We created detached context by iteratively added sentences to the two sides of the input sentence. The experimental design incorporated eight context window sizes ($k \in \{1,2,3,4,5,7,10,15\}$), creating contextual spans of $2k+1$ sentences (range: 3-31 sentences) when document boundaries permitted. For instance, a window size parameter $k=1$ produced a 3-sentence context window (preceding, target, and subsequent sentences). Each configuration underwent three repeated evaluation under identical experimental conditions to assess measurement stability.

Prediction alternations by context content

To investigate the relationship between the content of the context (i.e., History of Present Illness vs. Discharge Medications) and model behavior, we performed a differential prediction analysis on the i2b2 dataset. For this purpose, a pair of experiments was randomly selected from two settings: the baseline (no-context) setting and the section-level (detached-context) setting. For each input sentence, prediction alterations between these paired experiments were assessed and classified into three categories: (1) corrective flips, defined as instances where erroneous predictions in the baseline setting became correct when section-level context was introduced; (2)

degradation flips, defined as instances where initially correct predictions in the baseline setting became erroneous when section-level context was introduced; and (3) net improvements, calculated as the corrective flips minus degradation flips.

We also categorized the prediction alterations by sections. To quantify the correlation between section length and prediction alternations, we computed the Pearson correlation between the average number of words in a section category and the number of prediction alternations.

Results

Context evaluation

Baseline

Table 2 presents the performance of GPT-4o and the supervised BioClinicalBERT model on the MTSamples and i2b2 datasets when processing input sentences without context augmentation. BioClinicalBERT substantially outperformed GPT-4o in exact match metrics, with an F1 gap of approximately 0.2-0.3. This performance disparity narrowed considerably under relaxed match criteria, though GPT-4o still lagged by 0.02-0.1 F1 points.

Table 2. Performance of GPT-4o and supervised model (BioClinicalBERT) on MTSamples and i2b2 dataset when the input is given as a sentence without context.

Dataset	Model	P(exact)	R(exact)	F1(exact)	P(relax)	R(relax)	F1(relax)
MT Samples	GPT-4o	0.552±0.004	0.597±0.007	0.574±0.005	0.844±0.006	0.916±0.009	0.879±0.005
	BioClinical BERT [5]	0.785	0.785	0.785	0.915	0.887	0.901
i2b2	GPT-4o	0.480±0.005	0.445±0.003	0.462±0.004	0.864±0.003	0.797±0.006	0.829±0.003
	BioClinical BERT	0.78±0.009	0.811±0.009	0.795±0.005	0.919±0.006	0.956±0.017	0.937±0.009

Embedded context

As shown in Table 3, embedding structural context—either as section-level (MTSamples, i2b2) or note-level (i2b2) text—yielded minimal or detrimental effects on GPT-4o's performance. For MTSamples, section-level context has little impact of F1 score. It increased precision but also reduced recall. For i2b2, section-level context reduced F1 scores and note-level context reduced it further.

Table 3. Performance of GPT-4o on MTSamples and i2b2 dataset when the context is provided as embedded/detached section and note (note-level context only available for i2b2). Bolded values represent higher values in comparison to baseline (Sentence).

Dataset	Granularity	Integration	P(exact)	R(exact)	F1(exact)	P(relax)	R(relax)	F1(relax)
MTSamples	Sentence	N/A	0.552±0.004	0.597±0.007	0.574±0.005	0.844±0.006	0.916±0.009	**0.879±0.005**
	Section	Embedded	0.570±0.009	0.578±0.007	0.574±0.008	**0.873±0.016**	0.886±0.012	0.879±0.014
		Detached	**0.575±0.008**	**0.613±0.009**	**0.594±0.008**	0.850±0.007	**0.905±0.006**	0.877±0.004
i2b2	Sentence	N/A	0.480±0.005	0.445±0.003	0.462±0.004	**0.864±0.003**	0.797±0.006	0.829±0.003
	Section	Embedded	0.474±0.010	0.429±0.012	0.450±0.011	0.837±0.008	0.738±0.012	0.784±0.009
		Detached	**0.502±0.004**	**0.470±0.004**	**0.485±0.004**	0.862±0.004	**0.801±0.002**	**0.830±0.002**
	Note	Embedded	0.291±0.023	0.250±0.024	0.269±0.023	0.603±0.023	0.519±0.025	0.558±0.017
		Detached	0.469±0.004	0.421±0.004	0.444±0.003	0.842±0.003	0.750±0.003	0.793±0.002

Detached context

Table 4 also shows the performance of GPT-4o when section and note contexts were provided in a detached manner. Overall, MTSamples with detached section achieved the best exact match precision, recall, and F1 and the best relax match recall. i2b2 with detached section context achieved the best exact match precision, recall and F1, and best relax match recall and F1. Specifically, for MTSamples, detached section context improved exact match F1, precision and recall increasing by around 0.02. For i2b2, detached section context improved model performance higher, with exact match F1 and precision increasing by around 0.02 and recall increasing by around 0.04. Nevertheless, decreased model performance was observed for note-level context.

Context analysis

Model performance with varying context sizes

In Figure 2, we showed the performance of GPT-4o over the increasing number of sentences added as detached context (context created by adding sentences on both sides of the input sentence and context added as a prefix to the input sentence, e.g., 1 is 3 sentences in the detached context, 2 is 5 sentences, 3 is 7 sentences, etc.) in exact match F1 scores. Performance peaked when three sentences were added to each side of the input (total context size: 7 sentences, calculated as k×2+1, where k=3), achieving an F1 of 0.49. Beyond this threshold, performance declined monotonically, suggesting diminishing returns with excessive context. Notably, the optimal context size (k=3, 7 sentences) approximates the mean section length in the i2b2 dataset (5 sentences), indicating that effective context windows correspond to the typical length of sections in clinical documentation.

However, peak performance remained below that of the detached section-context configuration (F1=0.485 vs. 0.502), underscoring the superior utility of semantically bounded sections over arbitrary sentence windows. These results collectively indicate that while structural context remains optimal for precision tasks, a pragmatic balance—expanding context to include proximally relevant sentences—offers a computationally efficient alternative, achieving 95% of the maximal detached-section performance with minimal overhead.

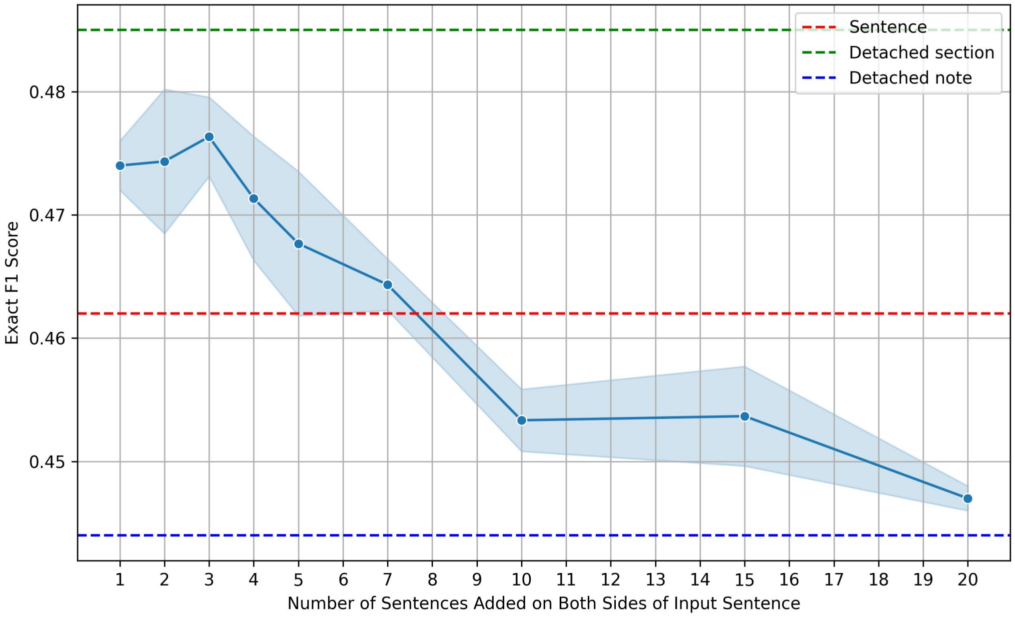


Figure 2. Exact match F1 scores were evaluated under varying extents of detached context, where 1, 2, 3, 4, 5, 7, 10, 15, or 20 sentences were symmetrically appended to both sides of the target sentence. Red/green/blue lines represented the model performance under the setting of no-context (sentence), detached section context and detached note context.

Prediction alternations by context content

Table 4 presents the analysis of prediction alternations when section-level context was introduced in a detached manner for GPT-4o on the i2b2 dataset. The prediction alternations were provided as three categories (1) corrective flips, defined as instances where erroneous predictions in the baseline setting became correct when section-level context was introduced; (2) degradation flips, defined as instances where initially correct predictions in the baseline setting became erroneous when section-level context was introduced; and (3) net improvements, calculated as the corrective flips minus degradation flips.

We observed that "Hospital Course", "Studies", and "Physical" section context exhibited the highest net improvements (Net: +24, +32, +36, respectively), while "History" showed no net change (Net=0). Pearson correlation analyses revealed statistically significant associations between section token length and prediction alterations: strong positive correlations were observed for corrective flips ($r=0.81$, $p<0.001$), degradation flips ($r=0.89$, $p<0.001$), and the total number flips ($r=0.85$, $p<0.001$), whereas the offset demonstrated a moderate correlation ($r=0.53$, $p<0.05$).

Table 4. Prediction alternation made by GPT-4o with detached section context on i2b2, in comparison to GPT-4o when no-context is provided, categorized by the type of sections. The total number of flips and section size (in tokens) are also shown. ρ represents the correlation. A value close to 1 represents high correlation and a value close to 0.5 represents weak correlation.

Section type	Corrective flips(ρ=0.81, p<0.001)	Degradation flips (ρ=0.89, p<0.001)	Net improvement (ρ=0.53, p<0.05)	Total number of Flips (ρ=0.85, p<0.001))	Section size (avg. # tokens)
Hospital Course	87	63	24	150	333
Studies	76	44	32	120	100
Physical	55	19	36	74	121
History	31	31	0	62	21
Unknown	18	8	10	26	101
Past Medical History	14	10	4	24	39
Discharge Diagnoses	7	4	3	11	16
Admission Diagnoses	6	2	4	8	18
Discharge Instructions	5	3	2	8	41
Discharge Medications	5	8	-3	13	103
Other Diagnoses	4	5	-1	9	32
Family History	3	0	3	3	15
Medications	3	3	0	6	44
Allergies	3	4	-1	7	10
Comments	2	0	2	2	20
Procedures	2	4	-2	6	16
Social History	1	0	1	1	32
Past Surgical History	1	0	1	1	29
Follow-up	1	0	1	1	14
Reason for Admission	1	0	1	1	18
Service	0	1	-1	1	8

Discussion and Conclusions

A critical finding is the divergence in efficacy between embedded and detached context methods. While embedded context—intuitively appealing for its seamless integration—resulted in performance degradation, detached context mechanisms consistently improved precision and recall. This degradation aligns with known LLM limitations in processing lengthy, unstructured inputs [15].

Detached context, in contrast, provides a clear advantage by separating contextual information from the primary extraction task. This approach limits its focus to the relevant input, resulting in improved precision, recall, and F1 scores. The section-level detached context, in particular, strikes an optimal balance, offering sufficient contextual information without overwhelming the model. Section boundaries carry outsized importance as they outperform contexts that are arbitrarily extended with padding sentences. The note-level context, however, may dilute relevance and introduce noise, as evidenced by the decline in performance.

When studying the impact of section content on prediction alternations, the results suggest that longer sections disproportionately amplify contextual reasoning opportunities, increasing both corrective and degradative flips. However, the weaker correlation for net improvement implies that greater context volume does not uniformly enhance net performance, as larger sections often introduce new errors. The findings underscore the dual role of structural context: while it expands the model's inferential scope, its utility remains contingent on section-specific relevance and coherence.

The observed performance gap between GPT-4o and supervised models (e.g., BioClinicalBERT) suggests that while LLMs hold promise, they are not yet a substitute for fine-tuned supervised methods in high-stakes clinical applications that require optimal accuracy. However, their zero-shot and few-shot capabilities make them a valuable tool for scenarios with limited annotated data.

In conclusion, this study provides a comprehensive analysis of the impact of context on LLM performance for clinical NER, revealing key insights into the trade-offs and benefits of different context incorporation strategies. The findings highlight the potential of detached context, particularly at the section level, to significantly enhance LLM performance without the drawbacks associated with embedded approaches. While supervised models remain superior in terms of exact match metrics, LLMs offer unique advantages in flexibility and scalability, making them an important complement in the clinical NER landscape. Future work should explore adaptive methods for context selection and integration, as well as the application of these findings to other clinical NLP tasks.

Acknowledgements

This work was supported by grants NIA R01AG080429 and RF1AG072799.

References

[1] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall, "2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text," *J. Am. Med. Inform. Assoc.*, vol. 18, no. 5, pp. 552–556, Sep. 2011, doi: 10.1136/amiajnl-2011-000203.

[2] E. Alsentzer *et al.*, "Publicly Available Clinical BERT Embeddings," in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, Minneapolis, Minnesota, USA: Association for Computational Linguistics, Jun. 2019, pp. 72–78. doi: 10.18653/v1/W19-1909.

[3] "Introducing Meta Llama 3: The most capable openly available LLM to date," Meta AI. Accessed: Jul. 13, 2024. [Online]. Available: https://ai.meta.com/blog/meta-llama-3/

[4] M. Li, H. Zhou, H. Yang, and R. Zhang, "RT: a Retrieving and Chain-of-Thought framework for few-shot medical named entity recognition," *J. Am. Med. Inform. Assoc.*, vol. 31, no. 9, pp. 1929–1938, Sep. 2024, doi: 10.1093/jamia/ocae095.

[5] Y. Hu *et al.*, "Improving large language models for clinical named entity recognition via prompt engineering," *J. Am. Med. Inform. Assoc.*, vol. 31, no. 9, pp. 1812–1820, Sep. 2024, doi: 10.1093/jamia/ocad259.

[6] "Hello GPT-4o | OpenAI." Accessed: Feb. 15, 2025. [Online]. Available: https://openai.com/index/hello-gpt-4o/

[7] X. Wang *et al.*, "Improving Named Entity Recognition by External Context Retrieving and Cooperative Learning," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds., Online: Association for Computational Linguistics, Aug. 2021, pp. 1800–1812. doi: 10.18653/v1/2021.acl-long.142.

[8] J. Lei, B. Tang, X. Lu, K. Gao, M. Jiang, and H. Xu, “A comprehensive study of named entity recognition in Chinese clinical text,” *J. Am. Med. Inform. Assoc. JAMIA*, vol. 21, no. 5, pp. 808–814, Oct. 2014, doi: 10.1136/amiajnl-2013-002381.

[9] Y. Luo, F. Xiao, and H. Zhao, “Hierarchical Contextualized Representation for Named Entity Recognition,” *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 05, Art. no. 05, Apr. 2020, doi: 10.1609/aaai.v34i05.6363.

[10] F. Chen, G. Zhang, S. Chen, T. Callahan, and C. Weng, “Clinical Note Structural Knowledge Improves Word Sense Disambiguation,” *AMIA Summits Transl. Sci. Proc.*, vol. 2024, pp. 515–524, May 2024.

[11] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[12] G. Team *et al.*, “Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context,” Dec. 16, 2024, *arXiv*: arXiv:2403.05530. doi: 10.48550/arXiv.2403.05530.

[13] M. Tepper, D. Capurro, F. Xia, L. Vanderwende, and M. Yetisgen-Yildiz, “Statistical Section Segmentation in Free-Text Clinical Records,” in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, Istanbul, Turkey: European Language Resources Association (ELRA), May 2012, pp. 2001–2008. Accessed: Oct. 31, 2022. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2012/pdf/1016_Paper.pdf

[14] Y. Vasiliev, *Natural Language Processing with Python and spaCy: A Practical Introduction*. No Starch Press, 2020.

[15] D. Reichenpfader, H. Müller, and K. Denecke, “A scoping review of large language model based approaches for information extraction from radiology reports,” *Npj Digit. Med.*, vol. 7, no. 1, pp. 1–12, Aug. 2024, doi: 10.1038/s41746-024-01219-0.

Appendices

1. GPT-4o prompt for baseline and embedded context studies

Task

Your task is to generate an HTML version of an input text, marking up specific entities related to healthcare. The entities to be identified are: 'medical problems', 'treatments', and 'tests'. Use HTML <span> tags to highlight these entities. Each <span> should have a class attribute indicating the type of the entity.

Entity Markup Guide

Use <span class="problem"> to denote a medical problem.

Use <span class="treatment"> to denote a treatment.

Use <span class="test"> to denote a test.

Leave the text as it is if no such entities are found.

Entity Definitions

Medical Problems are defined as: phrases that contain observations made by patients or clinicians about the patient’s body or mind that are thought to be abnormal or caused by a disease. They are loosely based on the UMLS semantic types of pathologic functions, disease or syndrome, mental or behavioral dysfunction, cellormolecular dysfunction, congenital abnormality, acquired abnormality,injury or poisoning, anatomic abnormality, neoplasticprocess, virus/bacterium, sign or symptom, but are not limited by UMLScoverage.

Treatments are defined as: phrases that describe procedures, interventions, and substances given to a patient in an effort to resolve a medical problem. They are loosely based on the UMLS semantic types therapeutic or preventive procedure, medical device, steroid, pharmacologic substance, biomedical or dental material, antibiotic, clinical drug, and drug delivery device. Other concepts that are treatments but that may not be found in UMLS are also included. Treatments that a patient had, will have, may have in the future, or are explicitly mentioned that the patient will not have are all marked as treatments.

Tests are defined as: phrases that describe procedures, panels, and measures that are done to a patient or a body fluid or sample in order to discover, rule out, or find more information about a medical problem. They are loosely based on the UMLS semantic types laboratory procedure, diagnostic procedure, but also include instances not covered by UMLS.

Annotation Guidelines

Only complete noun phrases (NPs) and adjective phrases (APs) should be marked. Terms that fit concept semantic rules, but that are only used as modifiers in a noun phrase should not be marked.

Include all modifiers with concepts when they appear in the same phrase except for assertion modifiers.

You can include up to one prepositional phrase (PP) following a markable concept if the PP does not contain a markable concept and either indicates an organ/body part or can be rearranged to eliminate the PP (we later call this the PP test).

Include articles and possessives.

Conjunctions and other syntax that denote lists should be included if they occur within the modifiers or are connected by a common set of modifiers. If the portions of the list are otherwise independent, they should not be included. Similarly, when concepts are mentioned in more than one way in the same noun phrase (such as the definition of an acronym or where a generic and a brand name of a drug are used together), the concepts should be marked together.

Concepts should be mentioned in relation to the patient or someone else in the note. Section headers that provide formatting, but that are not specific to a person are not marked.

Vital signs or vital signs with abnormal readings should be annotated as tests.

Medical specialists, services, or healthcare facilities should not be annotated, even if they might seem to fit into the categories of 'tests', 'treatments', or 'medical problems'. These entities are part of the healthcare delivery system and do not directly denote a test, treatment, or medical problem.

Consultation procedures should not be considered as tests.

Examples

Example Input1: At the time of admission , he denied fever , diaphoresis , nausea , chest pain or other systemic symptoms .

Example Output1: At the time of admission , he denied <span class="problem">fever</span> , <span class="problem">diaphoresis</span> , <span class="problem">nausea</span> , <span class="problem">chest pain</span> or other systemic symptoms .

Example Input2: He had been diagnosed with osteoarthritis of the knees and had undergone arthroscopy years prior to admission .

Example Output2: He had been diagnosed with <span class="problem">osteoarthritis of the knees</span> and had undergone <span class="test">arthroscopy</span> years prior to admission .

Example Input3: After the patient was seen in the office on August 10 , she persisted with high fevers and was admitted on August 11 to Cottonwood Hospital .

Example Output3: After the patient was seen in the office on August 10 , she persisted with <span class="problem">high fevers</span> and was admitted on August 11 to Cottonwood Hospital .

Example Input4: HISTORY OF PRESENT ILLNESS : The patient is an 85 - year - old male who was brought in by EMS with a complaint of a decreased level of consciousness .

Example Output4: HISTORY OF PRESENT ILLNESS : The patient is an 85 - year - old male who was brought in by EMS with a complaint of <span class="problem">a decreased level of consciousness</span> .

Example Input5: Her lisinopril was increased to 40 mg daily .

Example Output5: <span class="treatment">Her lisinopril</span> was increased to 40 mg daily .

Input Text: {input to extract entities from}

Output Text:

2. GPT-4o prompt for detached context studies

Task

Your task is to generate an HTML version of an input text, marking up specific entities related to healthcare. The entities to be identified are: 'medical problems', 'treatments', and 'tests'. Use HTML <span> tags to highlight these entities. Each <span> should have a class attribute indicating the type of the entity.

Entity Markup Guide

Use <span class="problem"> to denote a medical problem.

Use <span class="treatment"> to denote a treatment.

Use <span class="test"> to denote a test.

Leave the text as it is if no such entities are found.

Entity Definitions

Medical Problems are defined as: phrases that contain observations made by patients or clinicians about the patient’s body or mind that are thought to be abnormal or caused by a disease. They are loosely based on the UMLS semantic types of pathologic functions, disease or syndrome, mental or behavioral dysfunction, cellormolecular dysfunction, congenital abnormality, acquired abnormality,injury or poisoning, anatomic abnormality, neoplasticprocess, virus/bacterium, sign or symptom, but are not limited by UMLScoverage.

Treatments are defined as: phrases that describe procedures, interventions, and substances given to a patient in an effort to resolve a medical problem. They are loosely based on the UMLS semantic types therapeutic or preventive procedure, medical device, steroid, pharmacologic substance, biomedical or dental material, antibiotic, clinical drug, and drug delivery device. Other concepts that are treatments but that may not be found in UMLS are also included. Treatments that a patient had, will have, may have in the future, or are explicitly mentioned that the patient will not have are all marked as treatments.

Tests are defined as: phrases that describe procedures, panels, and measures that are done to a patient or a body fluid or sample in order to discover, rule out, or find more information about a medical problem. They are loosely based on the UMLS semantic types laboratory procedure, diagnostic procedure, but also include instances not covered by UMLS.

Annotation Guidelines

Only complete noun phrases (NPs) and adjective phrases (APs) should be marked. Terms that fit concept semantic rules, but that are only used as modifiers in a noun phrase should not be marked.

Include all modifiers with concepts when they appear in the same phrase except for assertion modifiers.

You can include up to one prepositional phrase (PP) following a markable concept if the PP does not contain a markable concept and either indicates an organ/body part or can be rearranged to eliminate the PP (we later call this the PP test).

Include articles and possessives.

Conjunctions and other syntax that denote lists should be included if they occur within the modifiers or are connected by a common set of modifiers. If the portions of the list are otherwise independent, they should not be included. Similarly, when concepts are mentioned in more than one way in the same noun phrase (such as the definition of an acronym or where a generic and a brand name of a drug are used together), the concepts should be marked together.

Concepts should be mentioned in relation to the patient or someone else in the note. Section headers that provide formatting, but that are not specific to a person are not marked.

Vital signs or vital signs with abnormal readings should be annotated as tests.

Medical specialists, services, or healthcare facilities should not be annotated, even if they might seem to fit into the categories of 'tests', 'treatments', or 'medical problems'. These entities are part of the healthcare delivery system and do not directly denote a test, treatment, or medical problem.

Consultation procedures should not be considered as tests.

Examples

Example Input1: At the time of admission , he denied fever , diaphoresis , nausea , chest pain or other systemic symptoms .

Example Output1: At the time of admission , he denied <span class="problem">fever</span> , <span class="problem">diaphoresis</span> , <span class="problem">nausea</span> , <span class="problem">chest pain</span> or other systemic symptoms .

Example Input2: He had been diagnosed with osteoarthritis of the knees and had undergone arthroscopy years prior to admission .

Example Output2: He had been diagnosed with <span class="problem">osteoarthritis of the knees</span> and had undergone <span class="test">arthroscopy</span> years prior to admission .

Example Input3: After the patient was seen in the office on August 10 , she persisted with high fevers and was admitted on August 11 to Cottonwood Hospital .

Example Output3: After the patient was seen in the office on August 10 , she persisted with <span class="problem">high fevers</span> and was admitted on August 11 to Cottonwood Hospital .

Example Input4: HISTORY OF PRESENT ILLNESS : The patient is an 85 - year - old male who was brought in by EMS with a complaint of a decreased level of consciousness .

Example Output4: HISTORY OF PRESENT ILLNESS : The patient is an 85 - year - old male who was brought in by EMS with a complaint of <span class="problem">a decreased level of consciousness</span> .

Example Input5: Her lisinopril was increased to 40 mg daily .

Example Output5: <span class="treatment">Her lisinopril</span> was increased to 40 mg daily .

Context

Here is the context for the following input text:

{ context in section/note}

Input Text: {input to extract entities from}

Output Text: