Large Language Models Provide Human-Level Medical Text Snippet Labeling

Ibtihel Amara[1,2*], **Haiyang Yu**[2], **Fan Zhang**[2], **Yuchen Liu**[2],
Benny Li[2], **Chang Liu**[2], **Rupesh Kartha**[2], and **Akshay Goel**[2]
[1] McGill University and [2] Google Research

Abstract

This study evaluates the proficiency of Large Language Models (LLMs) in accurately labeling clinical document excerpts. Our focus is on the assignment of potential or confirmed diagnoses and medical procedures to snippets of medical text sourced from unstructured clinical patient records. We explore how the performance of LLMs compare against human annotators in classifying these excerpts. Employing a few-shot, chain-of-thought prompting approach with the MIMIC-III dataset, Med-PaLM 2 showcases annotation accuracy comparable to human annotators, achieving a notable precision rate of approximately 92% relative to the gold standard labels established by human experts.

1 Introduction

Advanced natural language processing (NLP) tools especially generative language models have recently made a big difference in healthcare (Liu et al., 2023; Hu et al., 2023; Singhal et al., 2023; Goel et al., 2023; Tu et al., 2024). One key way NLP is used is to find important medical details, like diagnoses, within a patient's unstructured data. Clinicians can quickly search for medical conditions in these documents, speeding up their understanding of a patient's medical history.

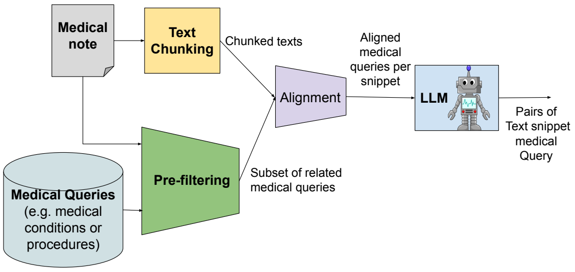
Medical note
Text Chunking
Chunked texts
Aligned medical queries per snippet
Alignment
LLM
Pairs of Text snippet medical Query
Medical Queries (e.g. medical conditions or procedures)
Pre-filtering
Subset of related medical queries

Figure 1: **Labeling Framework.** This consists of four components: (1) **Pre-filtering**; the query list is pre-filtered using a keyword search algorithm. (2) **Text Chunking**; the medical note is divided into smaller text snippets. (3) **Alignment**; The remaining queries are associated with the most relevant text snippets. (4) **LLM Labeling**; the text snippets and queries are sent to a large language model (LLM). The LLM confirms which conditions are truly relevant for each snippet.

In this work, we focus on identifying both potential and confirmed medical conditions throughout the various text snippets of information found in patients' medical records. Particularly, we establish a *mapping between a large comprehensive list of possible medical condition or procedures queries C and text snippets from clinical documents S.* We visualize the core task in Figure 4 in the Appendix. When establishing a connection between a medical condition or procedure and a snippet of medical text, we do not expect the text to include "supporting" components that are directly related to the condition or procedure. Instead, we anticipate that the labeler (here LLM) recognizes significant medical patterns, medications, and symptoms that point to a potential diagnosis (i.e. medical condition) or medical procedure. A straightforward example of this is as follows:

Text Snippet: *"The patient has been taking metformin 2500mg a day since last year."*
Possible LLM Condition/Procedure Labeling: *Diabetes and Polycystic Ovary Syndrome (PCOS).*
The rationale behind this labeling is that metformin is a medication commonly used in various medical treatments. Mastering this labeling process contributes to building the foundation for powerful information retrieval, search and summarization systems, which has the potential to revolutionize medical search and ultimately improve healthcare workflow. We summarize our main contributions as follows: (1) We demonstrate that LLMs can be used to identify potential labels (i.e medical conditions or procedures) with medical snippets reducing reliance on human experts. (2) We propose a cost-effective and efficient labeling framework with LLMs, which accelerates the annotation process

[*] Work done while the author was a research intern at Google Research.

[*]Corresponding authors: ibtihel.amara@mail.mcgill.ca, yuhaiyang@google.com, and zhanfan@google.com

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 185–195
June 21, 2024 ©2024 Association for Computational Linguistics

by reducing expensive LLM calls while preserving high labeling quality.

2 Related Work

Our work aligns with the field of Named Entity Recognition (NER) (Doan et al., 2012; Mullenbach et al., 2018; Yang et al., 2019; Goel et al., 2023; Guo et al., 2024; Ferraro et al., 2024). While NER primarily focuses on identifying and categorizing words into predefined entities such as procedure codes, medication codes, organizations, and others, our work takes a different approach. We adopt a unique methodology wherein we meticulously structure clinical documents by segmenting them into coherent and meaningful snippets. Our goal is to establish connections between these snippets and pertinent medical conditions or procedures drawn from a comprehensive list of medical queries. This approach allows us to not only identify potential medical conditions or procedures but also understand the context within the document, which ultimately will be useful for building and training medical search and retrieval systems.

3 Methodology

We provide in Figure 1 the general framework of our proposed labeling pipeline.

Pre-filtering. The first step in the pipeline involves pre-filtering a comprehensive list using cost-effective filtering strategies. This step aims to reduce the number of expensive calls to the LLM and avoid quality label loss (see Appendix I.1). There are several methods for implementing a pre-filtering step, such as embedding similarity, medical search engines, etc. We encourage researchers to explore other available and easy alternatives. In this work, we employed a keyword search algorithm. This technique expands the input queries (through query expansion) and looks for the matched text in the input document, which we regard as reference snippets. More details can be found in Appendix B.

Text Chunking. We broke down the patient's medical record into more manageable and informative text segments (i.e. medical snippets). We performed different chunking strategies (see Appendix D), and settled with a hybrid method involving a sentence-based (3-4 sentences) chunking algorithm with a constraint of 10-70 word tokens (Figure E).

Alignment. At this stage, we matched the remaining medical conditions and procedures to the corresponding text snippet. In particular, we opted for fuzzy matching. This can be considered as a secondary pre-filtering step at the snippet level. In our work, since our pre-filtering step outputs a reference snippet per condition or procedure, we attempted to locate these snippets within the different text chunks we have produced. This way, the condition becomes associated with the chunked snippet.[*]

LLM labeling. In the final stage of our framework, we paired the text snippets and their corresponding medical conditions. These pairs are then sent to the LLM using appropriate prompting strategies. The LLM assesses the relevance of the text snippet and medical condition in each pair. If it determines a condition to be relevant, the condition label is included as one of the final labels for that snippet.

4 Experimental Setup

Dataset and Pre-processing. We used the publicly available de-identified dataset MIMIC-III (Johnson et al., 2016). It is a collection of de-identified medical records and notes of more than 40,000 critical care patients at a large tertiary care hospital. It contains over two million unstructured clinical documents from nurses, physicians, etc. In our work, we randomly sampled 1000 patients and fetched all of their corresponding clinical records. Our pre-processing of the dataset was kept simplistic. We used simple regular expressions to identify formatting inconsistencies, such as extra spaces or tabs, in the clinical documents. We provide basic statistics in Section F about the sampled subset from the MIMIC-III dataset.

Human Labeling Workflow. The human labeling process was carried out in three separate rounds. In each round, a different group of medical expert raters was recruited to evaluate a distinct set of medical text snippets paired with a condition. Overall, we had 14 different medical experts as human annotators: 3 experts on the first round, 5 on the second, and 6 on the third round of labeling. Specifically, the raters were given a set of multiple-choice options ("Relevant", "Irrelevant", and "Not sure") and were asked to answer the following question: "Is the following text snippet relevant to the following medical condition/procedure?". The raters were given a random sample of snippets. In total, we collected 14,470 labeled snippet-condition pairs.

[*] It is important to note that the inclusion of this component is contingent upon the pre-filtering strategy that is ultimately adopted.

	Zero-shot						Zero-shot CoT						Few-shot CoT					
	NE			WE			NE			WE			NE			WE		
LLM Architecture	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
PaLM 2*	**83.55**	96.55	**89.58**	**71.79**	93.37	**81.17**	50.49	**91.07**	64.97	39.82	**87.25**	54.68	83.75	**96.28**	89.58	73.96	**94.25**	82.88
Med-PaLM 2	81.10	**98.51**	88.95	67.37	**96.92**	79.49	**91.74**	84.37	**87.91**	**81.49**	76.48	**78.91**	**92.70**	87.62	**90.09**	**84.94**	82.98	**83.95**

Table 1: **LLM performance compared to the golden human labels.** Med-PaLM 2 has the highest performance overall. NE: golden labels without rater exclusion; WE: golden labels with rater exclusion. P: precision; R: recall; F1: F1 score. *This is a fine-tuned PaLM 2 model variant for programming tasks.

LLM Labeling and Prompt Engineering Strategies. In this study we investigated LLM capabilities using simple prompt engineering techniques to more complex reasoning prompting strategies. We used zero-shot, few-shot, chain-of-thought (CoT) (Wei et al., 2022) , self-consistency CoT (Wang et al., 2022), and chain of verification (CoVe) (Dhuliawala et al., 2023). We assess these strategies on providing accurate labeling on medical snippets with respect to the “golden” labels obtained from human annotators. As for the LLM architectures, we used two different models: PaLM 2 (Anil et al., 2023) and Med-PaLM 2 (Singhal et al., 2023).

5 Results

We provide details about the basic statistics on both human labeled data and the sampled data from MIMIC-III in Appendix F.

Agreement Between Human Raters. Before relying on human labels, it is essential to assess their reliability and validity, especially when there is no clear or accessible ground truth label. To do this, we start by plotting the response distribution of each rater at each labeling round. Figure 2 exhibits significant variations within the different raters' responses. In round 1, for instance, two raters (raters 1 and 2) demonstrated a tendency to provide answers skewed towards the "irrelevant" category. In contrast, rater 3 maintained a balanced approach, assigning an equal number of responses to both the "irrelevant" and "relevant" categories. During the second round of the labeling process, raters 5, 6, and 8 exhibited a similar pattern of providing more "irrelevant" labels. In contrast, raters 4 and 7 produced more "relevant" responses. In the third round, we observe a similar distribution trend, which is predominantly characterized by a skew towards the "irrelevant" side. In Figure 3, we assess inter-rater reliability using Cohen's Kappa statistics (Viera et al., 2005; McHugh, 2012) and we provide in Appendix H the agreement interpretations. We observe that the level of agreement between raters varies across different rounds. In round 1 of labeling, the agreement ranges from "fair" to "moderate," indicating a practical level of consensus. However, in rounds 2 and 3, substantial variations emerge. In round 2, raters 5 and 6 exhibit a stronger agreement compared to other raters. In the third round, we observe a notable agreement between raters 11 and 12 and a moderate agreement between raters 10 and 11.

Golden Labels. Based on these reliability and agreement results, we decide to create *two types of golden labels*: (1) Majority vote with no rater exclusion [NE] and (2) Majority vote with rater exclusion [WE]. Indeed, for the first case, we mainly consider all of the raters' responses. As for the second version of golden labels, we consider only the majority voting of rater responses that are at least in a fair agreement with each other. In this case, we consider the following raters in each of the rounds (i.e. all raters in round 1, raters 5, 6, and 8 in round 2, and raters 10, 11, and 12 for round 3). We also applied a rigorous majority voting strategy. This involved selecting cases where there was a clear and consistent consensus among the raters. For instance, for a particular snippet-condition pair, we designated the snippet as relevant (associating it with the condition) only if all raters agreed that the condition was pertinent to the snippet. In cases where raters disagreed, we deemed the condition as "not sure", and excluded it from the evaluation.

LLM Performance on the Aggregated raters' labels. In Table 1, we compare the performance of different LLMs using different prompting strategies. Overall, Med-PaLM 2 achieves the highest precision across the different LLM architectures for each prompting strategy. This is likely because Med-PaLM 2 is specifically trained on medical text, which allows it to provide more precise results. However, when considering the recall metric, PaLM 2 achieves highest recall values, albeit with lower precision. When building a dataset for training medical retrieval systems, it is well preferred to have a good balance between precision and recall. Among the various prompting techniques, we

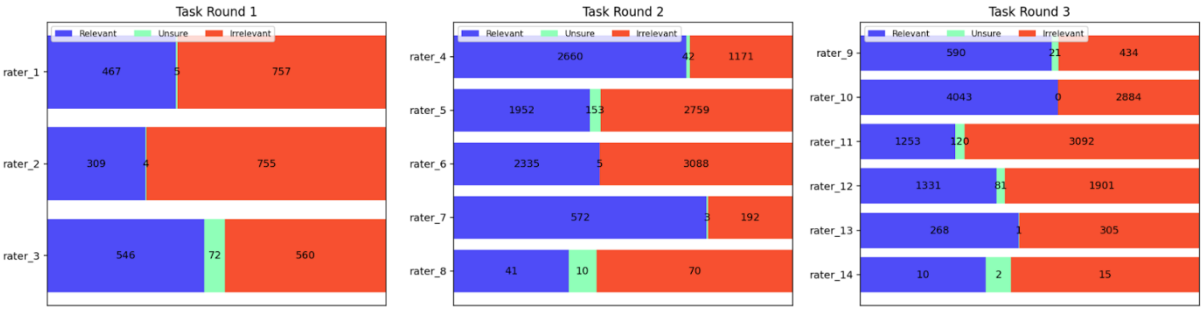

Figure 2: **Response Distribution of each Raters.** There are clear variations in the distribution of annotations across the raters.

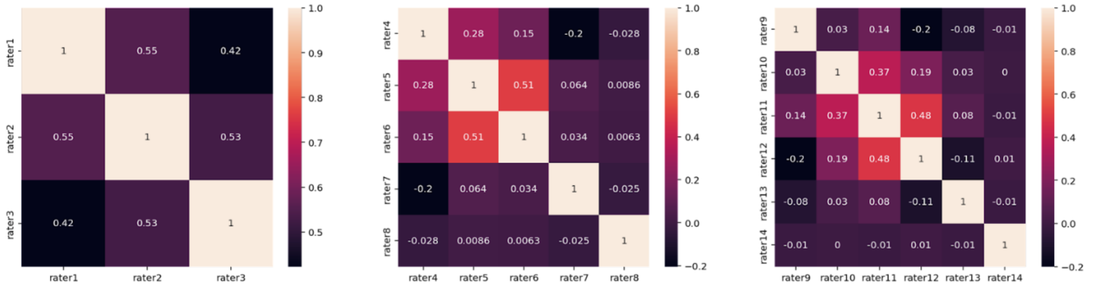

Figure 3: **Cohen Kappa's Inter-rater reliability.**

find that a few-shot CoT approach yields superior overall performance. Specifically, we observe improvements in both precision and recall metrics. Med-PaLM 2 outperformed in Few-shot CoT due to its medical focus. In zero-shot settings, PaLM 2 achieved top precision and F1 scores.

Beyond Basic Prompts. In addition to the three aforementioned prompts, we also explored two more prompting strategies and tested them on Med-PaLM 2: (1) self-consistency CoT and (2) Chain of Verification (CoVe). The accuracy of all these 5 prompts are shown in the Table 2. Note that the self-consistency prompting is based on the Few-shot CoT prompt with multiple runs using non-zero temperature (T=0.5). Although the ensemble result slightly outperforms the single run with T=0 (few-shot CoT), it requires multiple runs (three in our case), which substantially increases the time expenditure, hence we used the few-shot CoT for our final labeling task. Similarly, utilizing the CoVe prompt entails multiple rounds of verification to attain the final label. Each round demands distinct LLM invocations, rendering this method expensive.

Prompts	P	R	Acc.	F1
Zero-shot	78.73	97.30	89.97	87.03
Zero-shot CoT	92.79	72.59	88.57	81.45
Few-shot CoT	91.94	83.45	91.75	<u>87.49</u>
Self-Consistency	92.63	84.43	92.29	**88.34**
CoVe	73.98	77.67	82.83	75.78

Table 2: **Med-PaLM 2 performance on the NE dataset.** The highest F1 score is highlighted in bold, and the second-best score is underlined. Self-consistency yields the best performance. However, given that the few-shot prompt is less expensive than the self-consistency prompt, it is still a viable option.

Time Efficiency Comparison. On average, human raters took anywhere between 65 and 595 seconds (approximately 10 minutes) to review a single snippet, with an average time of 203 seconds. Considering an average of 8 conditions per snippet, this translates to roughly 24 seconds to review a snippet-condition pair. The latency of LLMs, on the other hand, varies depending on factors such as model architecture, size, inference infrastructure, and prompt strategies. However, on average, their latency is significantly lower than that of human raters.

6 Conclusion

We proposed a framework for labeling clinical notes. Our findings suggest that LLMs can produce high-quality medical data labels, which can serve as a valuable dataset for NLP tasks, such as information retrieval systems. These systems can help clinicians to be more efficient in their daily workflow by finding the key information faster and focus on pertinent facts within a clinical note.

7 Limitations

This work focused on a specific task: labeling medical conditions within clinical text snippets. While successful in this context, generalizing this approach to other scenarios might face limitations. Our keyword search method could miss relevant conditions not captured by the search algorithm. Additionally, the large language model (LLM) labeling is sensitive to the way it is prompted and requires further exploration to find optimal strategies for different use cases. Furthermore, the sentence-based chunking algorithm, while effective here, is specifically designed for the MIMIC-III dataset and may need adjustments for broader application. Finally, even human raters showed significant disagreement on labeling, highlighting the challenges posed by limited context in snippets and the inherent uncertainties within the medical domain, particularly when associating conditions with diverse symptoms. These limitations underscore the need for further research to improve generalizability and robustness when applying this type of system to broader medical text analysis tasks.

8 Ethical Statement

Labels created by LLMs might reflect biases inherent in the LLMs themselves. To some extent, these biases can be reduced by diversifying the LLMs, as this approach encourages the generation of more robust labels. However, even after implementing this strategy, biases may still persist. In the medical context specifically, additional alignment intervention methods can be utilized to modify the behavior of the LLM, presenting a potential solution to this challenge.

References

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Son Doan, Nigel Collier, Hua Xu, Pham Hoang Duy, and Tu Minh Phuong. 2012. Recognition of medication information from discharge summaries using ensembles of classifiers. *BMC medical informatics and decision making*, 12:1–10.

Antonino Ferraro, Antonio Galli, Valerio La Gatta, Mario Minocchi, Vincenzo Moscato, and Marco Postiglione. 2024. Few shot ner on augmented unstructured text from cardiology records. In *International Conference on Emerging Internet, Data & Web Technologies*, pages 1–12. Springer.

Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, et al. 2023. Llms accelerate annotation for medical information extraction. In *Machine Learning for Health (ML4H)*, pages 82–100. PMLR.

Yuting Guo, Yao Ge, and Abeed Sarker. 2024. Detection of medication mentions and medication change events in clinical notes using transformer-based models. *Studies in Health Technology and Informatics*, 310:685–689.

Yan Hu, Iqra Ameer, Xu Zuo, Xueqing Peng, Yujia Zhou, Zehan Li, Yiming Li, Jianfu Li, Xiaoqian Jiang, and Hua Xu. 2023. Zero-shot clinical entity recognition using chatgpt. *arXiv preprint arXiv:2303.16416*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Liwei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Zhengliang Liu, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Wei Liu, Dinggang Shen, Quanzheng Li, et al. 2023. Deid-gpt: Zero-shot medical text de-identification by gpt-4. *arXiv preprint arXiv:2303.11032*.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. *arXiv preprint arXiv:1802.05695*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, et al. 2024. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654*.

Anthony J Viera, Joanne M Garrett, et al. 2005. Understanding interobserver agreement: the kappa statistic. *Fam med*, 37(5):360–363.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Xi Yang, Jiang Bian, Yan Gong, William R Hogan, and Yonghui Wu. 2019. Madex: a system for detecting medications, adverse drug events, and their relations from clinical notes. *Drug safety*, 42:123–133.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, et al. 2024. Bench: Extending long context evaluation beyond 100k tokens. *arXiv preprint arXiv:2402.13718*.

A Visualization of Medical Note Labeling Task.

The goal is to categorize and classify each medical note text snippet into potential conditions. This pairing of text snippets and conditions can be highly valuable for training dense retrieval systems for medical notes pertaining to specific patients.

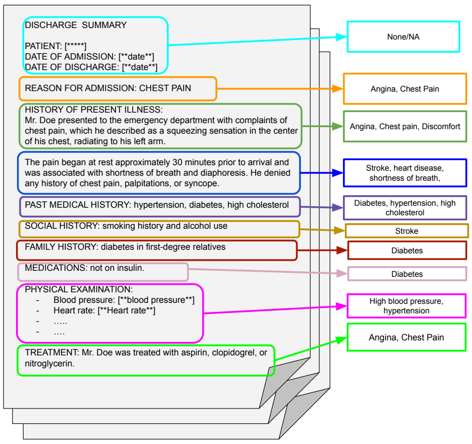
DISCHARGE SUMMARY
PATIENT: [*****]
DATE OF ADMISSION: [**date**]
DATE OF DISCHARGE: [**date**]
REASON FOR ADMISSION: CHEST PAIN
HISTORY OF PRESENT ILLNESS:
Mr. Doe presented to the emergency department with complaints of chest pain, which he described as a squeezing sensation in the center of his chest, radiating to his left arm.
The pain began at rest approximately 30 minutes prior to arrival and was associated with shortness of breath and diaphoresis. He denied any history of chest pain, palpitations, or syncope.
PAST MEDICAL HISTORY: hypertension, diabetes, high cholesterol
SOCIAL HISTORY: smoking history and alcohol use
FAMILY HISTORY: diabetes in first-degree relatives
MEDICATIONS: not on insulin.
PHYSICAL EXAMINATION:
- Blood pressure: [**blood pressure**]
- Heart rate: [**Heart rate**]
- .....
- ....
TREATMENT: Mr. Doe was treated with aspirin, clopidogrel, or nitroglycerin.
None/NA
Angina, Chest Pain
Angina, Chest pain, Discomfort
Stroke, heart disease, shortness of breath,
Diabetes, hypertension, high cholesterol
Stroke
Diabetes
Diabetes
High blood pressure, hypertension
Angina, Chest Pain

Figure 4: Medical Note Labeling Task

B An Example of Input and Output of the Implemented Mixer Search Algorithm for a Medical Note.

We used a keyword mixer search algorithm. This technique expands the input queries (via query expansion) and identifies their connections and locations within the input document. By positioning the keywords in the input document (25 tokens as the context with the searched keyword in the center), the algorithm generates reference sentences. Ultimately, the most representative reference sentence is given in relation to the input query (i.e. medical conditions/procedures). We illustrate the behavior of the mixer search algorithm as a technique for pre-filtering unlikely conditions from a medical note. Given a single query condition and the patient's clinical note, the algorithm identifies the most relevant text snippet from the document that is likely to be associated with the condition.

Input:
query: "coughing"
note: (note_id, the medical text)
Output: reference text from the medical note
"... Secretions: produced bloody and yellowish sputum with productive **cough** which was cleared with Yankauer and tracheal suction. Also of note ..."

C Identifying Sentence Boundaries

To identify sentence boundaries in the medical notes within MIMIC-III, we use regular expressions after some simple pre-processing as described in the experimental setup section. Regular expressions provide a flexible and efficient way to capture full sentences. They allow us to define patterns that match specific sentence-ending punctuation marks, such as periods (.), exclamation marks (!), and question marks (?). Additionally, regular expressions can be used to handle more complex cases, such as sentences that end with abbreviations or quotations.

D Note Chunking/Segmentation Strategies.

We implemented several ways of text chunking to split each medical note properly:
(1) Sentence-base (SB) segmentation: The medical note is fragmented according to a collection of one or more sentences. We divide the document into n non-overlapping sentences without regard for the notes' structure and sectioning.
(2) Word-base (WB) segmentation: The medical note is fragmented according to a collection of one or more word tokens. One thing to note is that this word base does not consider the cut offs. In other words, it would take the number of words given in the input regardless of it being an incomplete or full sentence. For our use case, snippets would be more readable (contextually and grammatically correct) for later human and LLM labeling.
(3) Sentence-word fusion (SWF) segmentation: One major thing we noticed during the execution of our algorithms is that we were getting a lot of very short sentences. To mitigate this, we implemented a hybrid version of the snipping algorithms above. We considered a sentence-based text segmentation, with a constraint on the number of words admissible for each segment via a range threshold. In this work, we chose a balance of 3-4 sentences with the constraint of 10- 70 word tokens.

E Distribution of the Number of Tokens for Different Chunking Algorithms.

The MIMIC-III dataset was used to extract medical notes, and the chunking algorithm was then used to obtain the distribution of token counts. Naively chunking (into four sentences) resulted in very short sentences, mainly due to the formatting of MIMIC-III and the simple pre-processing done on

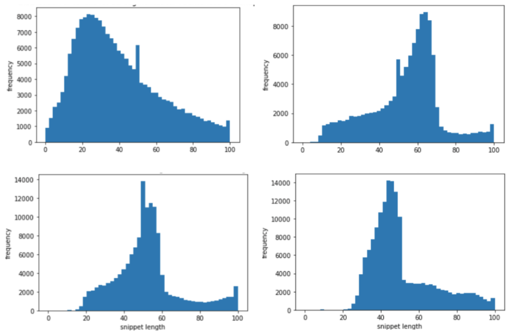

Figure 5: **Distribution of token count using different chunking strategies.** Top left: Sentence-based segmentation (4 sentences per snippet). Top right: Sentence-based segmentation with a token count constraint of 10-70. Bottom left: Sentence-based segmentation with a token count of 20-60. Bottom right: Sentence-based segmentation with a token count constraint of 30-50.

the notes. To address this, we opted for a sentence-based constraint on token count, resulting in improved snippets. A 10-70 constraint was chosen as it captures an appropriate amount of atomical (singleton) information, while larger constraints could lead to more extensive snippets with more information.

F Statistics on Human Labeled Data

There are totally 46,146 patients and 2,083,159 notes in the MIMIC3 dataset. We collected 499 medical conditions as queries and sampled 1,000 patients randomly to generate the labeled data for future model training. For the evaluation purpose, we launched three runs of human evaluation: the first run randomly sampled 100 chunked note snippets across patients and notes, the second and third runs sampled 5 patients each and totally 338 notes and 1,048 note snippets. We asked at least three medical expertise to evaluate the data independently in each human evaluation run, and at the end we had 14 independent raters working on 1,079 note snippets and 14,470 snippet-condition pairs. Due to the raters' availability, 9,812 of the snippet-condition pairs were evaluated by three raters independently, 896 of them were evaluated by two raters, and the left 3,762 pairs were evaluated by only one rater.

The basic statistics of the note snippets and condition queries are shown in Figure 3. Because of the settings of our chunking algorithm, most of the snippets have reasonable length (around 60 tokens). Most of the condition queries are single words or short phrases with 2 to 3 tokens. The keyword mixer search algorithm efficiently narrows the conditions for each snippet: on average, each snippet has about 13 relevant conditions (compare with the full list of 499 conditions), which will largely reduce the time cost of LLM labeling. About half of these pre-filtered snippet-condition pairs were further labeled as true relevant pairs, according to the majority voting of human raters.

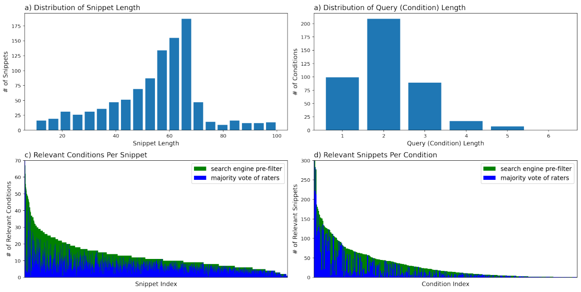

Figure 6: **Basic Statistics of the Note Snippets and Condition Queries.** a) Distribution of snippets over length (token counts); b) Distribution of condition queries over length (token counts); c) Counts of relevant conditions of each snippet (green: based on the search engine pre-filter results; blue: majority voting from human raters); d) Counts of relevant snippets of each condition.

G Prompting strategies

G.1 Zero shot

You are an expert medical assistant. Your task is to give an answer of Yes/No for the relevance between a snippet and condition pair. A snippet is relevant to a condition if it includes information about the symptoms, assessments, labs, vitals, medications, procedures, or past medical history of a patient that is relevant to the given condition.

G.2 Zero-shot CoT

You are a clinical specialist. You will be given a medical note snippet (S) and a medical condition or procedure (C). Your task is to mark whether the snippet S mentions meaningful information for C to you. Mark the answer with a binary number (0 or 1). A score of 0 indicates that the snippet does not contain meaningful content to the condition, while a score of 1 indicates that the snippet contains meaningful content. Walk me through your thoughts.

If C is a condition, snippet S contains meaningful information for C if it satisfies one of the following criterias:

(1) The snippet contains description of the condition (including explicit denial of the condition).
(2) The snippet contains description of a common cause to the condition.
(3) The snippet contains description of symptom(s) that are strongly correlated with the condition.
(4) The snippet contains description of findings that could suggest the condition (including findings that can rule out this condition).

If C is a procedure, snippet S contains meaningful information for C if it contains description of the procedure C.
S: snippet.
C: condition.
A:

G.3 Few-shot CoT

You are an experienced clinician. You will be given a medical note snippet (S) and a medical condition or procedure (C).
Your task is to decide whether the snippet mentions useful information to a clinician for understanding the condition or procedure.
Think step by step without hallucination and provide a final Yes/No answer.

If C is a [condition], snippet S contains useful information for C if it satisfies one of the following criteria:
(1) The snippet contains information that clearly certifies or excludes C.
(2) The snippet contains highly specific information for C (symptoms, signs, or test values).

If C is a [procedure], snippet S contains useful information for C if it contains one of the following criteria.
(1) The snippet contains information that clearly certifies or excludes C.
(2) The snippet mentions clinical conditions that are highly specific to C.

Example1: C: foot pain S: ros: the patient denies any fevers, chills, weight change, nausea, vomiting, abdominal pain, diarrhea, constipation, melena, hematochezia, chest pain, shortness of breath, orthopnea, pnd, lower extremity edema, cough, urinary frequency, urgency, dysuria, lightheadedness, gait unsteadiness, focal weakness, vision changes, headache, rash, or skin changes. A: Step 1. C (foot pain) is a common [condition] that refers to pain in the foot (lower extremity).
Step 2. Is there an explicit positive/negative signal of C in S? : No, S contains multiple negative symptoms as part of a ROS but does not contain any features related to foot pain.
Thus the answer is No.

Example2
....
ExampleN

C: condition
S: snippet
A: """

G.4 Chain-of-Verification CoVe

BASELINE PROMPT = You are a medical specialist/clinician. You will be given a medical note snippet (S) and a condition/procedure (C).
Your task is to answer the below question (Q) correctly and concisely with a Yes/No answer then provide your explanation and thoughts.
Q: Does the snippet (S) directly or indirectly relate to the condition or procedure (C)?
A direct relationship is when the snippet (S) contains a description of the condition/procedure (C) or perhaps a common cause to the condition/procedure (C).
An indirect relationship is when the snippet (S) contains description of symptoms that are strongly correlated with the condition/procedure (C) or findings that could suggest the condition/procedure (C). Provide clear step by step explanations and thoughts.
S: snippet
C: condition
Answer:
VERIFICATION QUESTIONS = You are a medical expert. You will be given a medical note snippet (S), a condition (C), a question (Q), and a baseline response (BR) coming from another clinician.
Your goal is to generate three verification questions that relate to both (S) and (C). These verification questions should give a clearer guidance on how to get factual answers based on the (Q) and (BR). They are meant for verifying the factual accuracy in the baseline response (BR). The verification questions must show consistency with (Q), (BR), (S), and (C).
S: snippet

C: condition
Q: Does the snippet (S) directly or indirectly relate to the condition or procedure (C)?
BR:baseline response
Verification Questions:

EXECUTE PLAN PROMPT = You are a medical expert. You will be given a medical note snippet (S), a condition (C), some verification questions (VQ) to answer as a second opinion expert.
Your task is to provide answers to the verification questions (VQ) as correctly as possible based on the given snippet (S) and condition (C). The verification questions (VQ) could be tricky as well, so think step by step and answer them correctly.
S: snippet
C: condition
VQ: verification questions
Answer:

REFINEMENT = You are a medical expert. You will be given a medical note snippet (S), a condition (C), a medical question (MQ), a baseline response (BR), some verification questions (VQ) related to all the above, and their corresponding verification answers (VA) provided by another medical assistant.
S: snippet
C: condition
MQ: Does the snippet (S) directly or indirectly relate to the condition or procedure (C)? A direct relationship is when the snippet (S) contains a description of the condition/procedure (C) or perhaps a common cause to the condition/procedure (C).
An indirect relationship is when the snippet (S) contains description of symptoms that are strongly correlated with the condition/procedure (C) or findings that could suggest the condition/procedure (C).
BR: baseline response
VQ: verification questions
VA: verification answer
Your task is to analyze all of the above information and provide a refined [Yes/No] answer to the medical question (MQ). You must answer with a [Yes/No] response.
Make sure to provide clear explanations, a good walk through of your thoughts based on the information in (S), (C), (MQ), (BR), (VQ), and (VA).
Answer:

H Interpretation of Cohen Kappa's statistics.

Table 3 provides detailed breakdown for interpreting the cohen Kappa value.

Kappa Values	Agreement
<0	Less than chance agreement
0.01 - 0.20	Slight agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 0.99	Almost perfect agreement

Table 3: Interpretation of Kappa statistics (Viera et al., 2005)

I Frequently Asked Questions

I.1 Why was the LLM not given a list of medical conditions to choose from when labeling a medical text snippet?

Research has shown that LLM performance is correlated with the number of tokens provided in the context (Zhang et al., 2024). Therefore, it is not sensible to use a voluminous and comprehensive list of medical conditions and provide it to the LLM for selection. An alternative and better strategy would be to provide the LLM with medical snippet-condition pairs and ask it to determine the relevance of each pair, which is the strategy used in this work.

Although this approach can reach high accuracy, it presents challenges too: as performing multiple inferences on the LLM can be computationally expensive and may result in long labeling times if resources are limited. For example, to label millions of snippets with associated thousands of conditions, the time complexity would be in the order of $O(10^8)$ or $O(10^9)$, and since LLM inference usually is slow (in seconds) thus the time cost will be in the order of $O(10^3)$ or $O(10^4)$ days. Thus, we need a fast condition filter before sending the data to LLM.

J The Comprehensive List of Conditions used in this study.

Our study considered the 499 most prevalent, frequently encountered and queried medical conditions and procedures in medical notes. While we only provide 20 examples below, more detailed information is available upon request: amputation, anemia, angioedema urticaria, angioplasty of blood vessel, burn, cardiac abscess, cardiac arrest, corneal disease, cough, covid 19, flank pain, foot pain, fracture, fracture fixation, insulin resistance, lung malignancy, ovarian abscess, ophthalmologic procedure, oropharyngeal infection, pancreatitis, etc