Zero-shot Large Language Models for Long Clinical Text Summarization with Temporal Reasoning

Maya Kruse[1], Shiyue Hu[1,2], Nicholas Derby[1,2], Yifu Wu[1], Samantha Stonbraker[1]
Bingsheng Yao[3], Dakuo Wang[3], Elizabeth Goldberg[1], Yanjun Gao[1]
[1]University of Colorado Anschutz Medical Campus
[2]University of Colorado Boulder
[3]Northeastern University

Correspondence: yanjun.gao@cuanschutz.edu

Abstract

Recent advancements in large language models (LLMs) have shown potential for transforming data processing in healthcare, particularly in understanding complex clinical narratives. This study evaluates the efficacy of zero-shot LLMs in summarizing long clinical texts that require temporal reasoning, a critical aspect for comprehensively capturing patient histories and treatment trajectories. We applied a series of advanced zero-shot LLMs to extensive clinical documents, assessing their ability to integrate and accurately reflect temporal dynamics without prior task-specific training. While the models efficiently identified key temporal events, they struggled with chronological coherence over prolonged narratives. The evaluation, combining quantitative and qualitative methods, highlights the strengths and limitations of zero-shot LLMs in clinical text summarization. The results suggest that while promising, zero-shot LLMs require further refinement to effectively support clinical decision-making processes, underscoring the need for enhanced model training approaches that better capture the nuances of temporal information in long context medical documents.

1 Introduction

Electronic Health Records (EHRs) encapsulate a wide range of multi-modal data, such as vital signs, laboratory results, radiology findings and free text clinical notes (Mohsen et al., 2022; Li et al., 2022). They are organized across various timestamps, reflecting the dynamic nature of patient care. Accurately summarizing this data is crucial, as it provides healthcare professionals with insights into patient progress and assists in clinical decision-making (Adams et al., 2021; Gao et al., 2023a; Laxmisan et al., 2012; Pivovarov and Elhadad, 2015; Liang et al., 2019). However, the complexity and volume of data within EHRs pose significant challenges for effective summarization, highlighting the need for sophisticated tools that can handle such detailed and varied information efficiently.

Recent advancements in large language models (LLMs) have demonstrated their proficiency in a variety of clinical NLP tasks, notably in replying to patient messages, summarizing patient diagnoses, and generating discharge summaries (Silcox et al., 2024; Wachter and Brynjolfsson, 2024). Despite these successes, most LLM applications do not fully account for the entire patient trajectory during their hospital stay, often simplifying or omitting intricate details crucial for comprehensive patient care. Traditional techniques like Retrieval-Augmented Generation (RAG) are designed to process long contexts (Gao et al., 2023b; Lewis et al., 2020). Existing methods examine RAG on clinical text for diagnosis prediction and discharge summarization (Myers et al., 2024; Lyu et al., 2024). Yet, their effectiveness in managing long clinical documents embedded with temporal data remains largely unexplored. This study aims to address this gap by evaluating how well current LLMs, including those enhanced with RAG, perform in summarizing long clinical narratives that incorporate essential temporal information.

In this paper, we specifically investigate two established tasks within the domain of clinical NLP: Discharge Summary and Progress Note Generation, but within a new setting for long documents that features multiple timestamps. We utilize a publicly available dataset, the Medical Information Mart for Intensive Care (MIMIC) dataset to conduct our investigations (Johnson et al., 2020). We investigated three state-of-the-art LLMs, Qwen2.5-7B (Yang et al., 2024), Mistral-7B-Instruct-v0.1 (Jiang et al., 2023) and Llama3-8B-Instruct (AI@Meta, 2024). Results indicate that current LLMs still struggle with the task of clinical text summarization for long context documents as well as temporal reasoning over patient trajectories. Even though RAG shows promise, the overall performance is still on

the lower end, highlighting the need for further work - we aim to investigate an event extraction approach for future research.

2 Related Work

Hospital course summarization is a topic that has been addressed by many previous works. At ACL 2024, a shared task on discharge summary generation (Xu et al., 2024) was included in the BioNLP workshop, focusing on the generation of the 'Brief Hospital Course' and 'Discharge Instructions' sections. This task is similar to ours, but does not include the 'Primary Diagnosis' section. In the paper 'SPEER: Sentence-Level Planning of Long Clinical Summaries via Embedded Entity Retrieval' (Adams et al., 2024), a novel approach to hospital course summarization is introduced, which performs a content selection step before summarization, thus boosting the coverage of salient information as well as faithfulness in the generated summaries. Like our paper, it also focuses on long document summarization. (Gao et al., 2022b) address a similar problem to progress note generation - taking a patient's progress notes as input, a list of problems in the patient's daily care plan is generated. Additionally, they experiment with data augmentation and domain adaptation pre-training.

3 Dataset and Tasks

3.1 Dataset

The main dataset utilized in this paper is MIMIC-III, which was chosen over MIMIC-IV due to its larger variety of free text note types - while MIMIC-IV only contains discharge summaries and radiology notes, MIMIC-III features both of these note types along with physician and nursing notes, ECG reports and others. In addition to the unstructured note data, the dataset also contains large amounts of structured, or tabular, data. There are many different types of structured data available, but we only considered the chart, lab and input events, as well as patient medications. Chart events are routine vital signs as well as additional information related to patient care such as ventilator settings, mental status etc taken during the patient's hospital stay. Lab events refer to values taken from laboratory measurements, such as white blood cell count, cholesterol levels etc. Many of the values listed in lab events are also found in chart events, leading to duplication of information. Input events are fluids administered to the patient, which could be oral/tube feedings or IV infusions, and medications are self-explanatory - they are any medications the patient receives. These distinct data types are synthesized into an admission chronology - all data recorded for a patient's admission is gathered and ordered by timestamp.

It is a well-known problem that clinical information is often recorded several times over, making it more difficult to extract the relevant information (Shoolin et al., 2013). Additionally, this extremely large amount of data can be a problem when working with LLMs, which have a limited context window for input data. In order to counteract this problem and reduce redundancy within the data, some minor preprocessing was done. First, lab events were filtered to only include values that had been flagged with a 'WARNING' - the idea being that values which fall outside of the normal range are more likely to represent salient clinical information. After this initial filtering, values present in both lab and chart events are removed from chart events, ensuring only one remains. In terms of target population, we focus on older patients (65 years and older) who have been admitted to the ICU. We chose this patient cohort because they typically exhibit the highest level of medical complexity, making their care scenarios particularly challenging to manage. This complexity often includes multiple chronic conditions and a higher frequency of hospital visits, which necessitates clear and concise medical summarization to ensure effective and streamlined care. Streamlined summarization of clinical data could lessen their burden and aid them in making more informed decisions. We further limit our selection to admissions whose length of stay is three days or longer - this is because we are investigating long document summarization as well as the ability of LLMs to reason over a progression of clinical events.

3.2 Tasks

3.2.1 Discharge Summary Generation

Discharge summaries are a crucial part of a patient's hospital care process, as it contains a comprehensive overview of their hospital stay and the relevant clinical events which took place during it. It is usually written by the physician after the patient is discharged, and contains three main sections:

1. the primary diagnosis: a list of diagnoses upon discharge

2. an overview of the hospital course: clinical summary of treatments and events during hospital stay

3. discharge instructions: this section contains a variety of instructions for the patient post-discharge, such as medication and dietary instructions or plans for therapy and medical follow-ups

Discharge Summary Generation as a task has been covered in several previous works, including (Xu et al., 2024) and (Ando et al., 2022). Many of these previous works focus on a particular section of the summary, such as the Brief Hospital Course (BHC) or discharge instruction section, but in this paper we attempt to generate all three sections.

We extract all structured and unstructured data pertaining to a specific hospital admission and concatenate it into one chronologically ordered document. Structured data is converted into a narrative format prior to concatenation. Since we are dealing with longer admissions (3 days or longer), a significant amount of data is accumulated, and we chose to focus on only the last 24 hours of EHR data in order to avoid overloading the LLM with data.

3.2.2 Progress Note Generation

During a patient's stay in a hospital, a written record of their treatment, clinical status and progress is kept. Progress notes are an important part of this record, containing key information about the patient's current condition and treatment, updated daily by the physician. The structure of a progress note typically follows the SOAP format, and consists of four sections: Subjective, Objective, Assessment and Plan. The subjective section consists of unstructured data in the form of free text, with descriptions of the patient's symptoms, status and treatment, while the objective section contains structured data such as lab results and other charted values. In the Assessment section, the patient's main symptoms and diagnoses are laid out, and in the Plan section these diagnoses and problems are each addressed with a detailed action or treatment plan. For this task of Progress Note Generation, we chose to focus only on the Assessment and Plan sections, as the content of these sections represents the reasoning over the Subjective and Objective sections, extracting salient information and developing the diagnoses and corresponding treatment plans from them (Gao et al., 2022a). In general, the task is always to generate a progress note for the current day when given the EHR chronology for that day, however we examine different methods of incorporating previous progress notes into the input data. As with the Discharge Summary Generation task, the input is a combination of structured and unstructured data, however progress notes are excluded from the EHR chronology to prevent the model from ingesting ground truth data.

- Baseline method (no prior note): NO previous progress notes are included in the input data
- Method 1 (Single-Day Context): the progress note from the previous day is included in the input data
- Method 2 (Multi-Day Context): progress notes from ALL previous days are included in the input data

It is important to note that we conduct evaluation on a daily level, so while methods 1 and 2 include ground truth data, it is not the ground truth data that is being evaluated against at the current time step. This means that for day i, the input data includes progress notes up to and including day i-1 (depending on the method employed), but the progress note for day i is held out and compared to the generated progress note for that day. This choice is made to mirror the real-life clinical workflow, as physicians often rely on previous progress notes when writing current ones. We also experiment with a setting in which the previous progress notes incorporated into the input data are those previously generated instead of the ground truth ones.

4 Methods

The tasks outlined above were tested on three models:

- Mistral-7B-Instruct-v0.1
- Llama3-8B-Instruct
- Qwen2.5-7B

Our experiments are designed to investigate several areas of interest: multimodality, temporal understanding and long contexts. The experiments created for the task of discharge summary generation attempt to address all three. First, to take a closer look at the way LLMs handle multimodality, we compare data that is strictly tabular, strictly unstructured (free text notes) and a combination

of both. Second, the temporally ordered data is compared with randomly shuffled data, to investigate the extent to which the model relies on the provided temporal information.

The experiments for the progress note generation task mainly focus on the long context aspect. The three methods have increasing context lengths, and thus represent different levels of long context. With longer contexts, more salient information is present, though it also becomes more difficult to extract it, since the amount of extraneous information increases as well. This trade-off is something the different methods were designed to address.

4.1 Approaches

Two approaches are contrasted on the Discharge Summary and Progress Note Generation tasks. These approaches are a baseline direct generation approach and a more structured RAG approach.

4.1.1 Direct Generation

This approach is extremely straightforward - the patient's chronology (in the corresponding format for each task) is simply fed into the model without any additional structure. In this way, the approach functions as a baseline of how current popular models are able to handle clinical data and tasks.

4.1.2 RAG

Given the massive amount of data that is accumulated during a patient's hospital stay, a RAG system is a natural choice. The data is organized into a database, and only the relevant sections (with regard to a query) are retrieved, making the model less prone to 'data overload'. For this setup, we chose the same selection as models used for direct generation.

Our choice of embedding was informed by (Myers et al., 2024), which identified BGE embeddings as the highest performing on a variety of models. Additionally, the queries presented in the paper were slightly modified and used to perform query optimization. Hyperparameter optimization was carried out on the standard RAG hyperparameters chunk size, chunk overlap and top-k retrieved documents.

4.2 Evaluation

Several metrics were chosen for these tasks. ROUGE-L (Lin, 2004) is a similarity metric, and the standard for evaluating summarization, that calculates the longest common subsequence. BERTScore (Zhang et al., 2019) is used for semantic similarity - we chose to utilize SapBERT embeddings (Liu et al., 2020), as they are optimized for the biomedical domain. (Croxford et al., 2024) showed that BERTScore using SapBERT has the strongest alignment with human evaluation among the metrics tested, which is why it is included here. Finally, we also calculate the CUI f-score - clinical entities (from the ULMS (Lindberg et al., 1993)) from the generated and ground truth files are extracted, and the f-score is calculated over them. The following semantic types were used for the respective sections of discharge summaries:

- Diagnosis: 'Disease or Syndrome' and 'Pharmacologic Substance'
- Brief Hospital Course: 'Pharmacologic Substance', 'Finding', 'Therapeutic or Preventative Procedure', 'Molecular Function' and 'Sign or Symptom'
- Discharge Instructions: 'Pharmacologic Substance', 'Health Care Activity', 'Finding', 'Therapeutic or Preventive Procedure' and 'Organic Chemical'

The semantic types used for Progress Note generation were 'Intellectual Product', 'Qualitative Concept', 'Idea or Concept', 'Functional Concept', 'Disease or Syndrome'.

5 Results

5.1 Direct Generation

5.1.1 Discharge Summary Generation

Among the models tested, Qwen2.5-7B performed the best, though the margin to the other models was not significant - they are usually within 2-3 percentage points of each other. Qwen2.5-7B's results are presented in Table 1. The 'note only' modality performs well across models, suggesting that the inclusion of tabular data is not very beneficial for this task. Overall, the 'Hospital Course' section achieves the highest scores, closely followed by 'Discharge Instructions' and then 'Primary Diagnosis'. Interestingly, there is no big difference between the shuffled and chronological tabular, showing that the temporal information present in the data does not play a critical role in influencing performance.

	Diagnosis	Hospital Course	Discharge Instructions
All	CUI: 6.26 ROUGE-L: 4.09 SapBERT: 49.92	CUI: 7.62 ROUGE-L: 12.70 SapBERT: 64.01	CUI: 5.56 ROUGE-L: 11.44 SapBERT: 64.40
Note only	**CUI: 6.99** **ROUGE-L: 5.95** **SapBERT: 54.12**	**CUI: 9.50** **ROUGE-L: 13.82** **SapBERT: 68.28**	**CUI: 5.80** **ROUGE-L: 12.01** **SapBERT: 67.14**
Tabular only	CUI: 5.89 ROUGE-L: 3.67 SapBERT: 46.63	CUI: 6.10 ROUGE-L: 11.03 SapBERT: 60.71	CUI: 5.63 ROUGE-L: 11.63 SapBERT: 63.15
Shuffled tabular	CUI: 3.63 ROUGE-L: 3.49 SapBERT: 49.13	CUI: 6.90 ROUGE-L: 11.89 SapBERT: 64.59	CUI: 4.65 ROUGE-L: 10.97 SapBERT: 64.93

Table 1: Results on the highest-performing model for Discharge Summary generation, Qwen2.5-7B

Setting	Method	CUI	ROUGE-L	SapBERT
GT	No Prior Note	20.14	8.58	67.37
	Single-Day Context	28.34	16.99	68.15
	Multi-Day Context	26.54	**15.78**	**67.98**
GEN	No Prior Note	20.14	8.58	67.37
	Single-Day Context	**36.11**	20.47	69.14
	Multi-Day Context	32.25	19.52	67.76

Table 2: Progress Note generation results on Llama3-8B-Instruct

Setting	Method	CUI	ROUGE-L	SapBERT
GT	No Prior Note	15.46	11.07	67.55
	Single-Day Context	25.61	13.80	71.32
	Multi-Day Context	**32.25**	**19.52**	67.76
GEN	No Prior Note	15.46	11.07	67.55
	Single-Day Context	30.16	14.50	**71.65**
	Multi-Day Context	25.24	14.54	70.64

Table 3: Progress Note generation results on Qwen2.5-7B

Model	Section	Prompt	CUI	ROUGE-L	SapBERT
Mistral	Diagnosis	4	30	3.95	54.83
	Hospital Course	3	3.36	12.36	67.32
	Discharge Instructions	3	15.74	16.22	65.03
Llama	Diagnosis	3	25.71	3.83	53.9
	Hospital Course	3	3.41	13.72	70.12
	Discharge Instructions	3	14.54	16.38	60.62
Qwen	Diagnosis	4	26.67	0.72	68.77
	Hospital Course	3	5.7	13.19	70.81
	Discharge Instructions	5	22.2	16.33	62.22

Table 4: Query optimization results for Discharge Summary generation

Method	GT	GEN
No Prior Note	Prompt 4 (Mistral) Prompt 3 (Llama) Prompt 4 (Qwen)	Prompt 4 (Mistral) Prompt 1 (Llama) Prompt 5 (Qwen)
Single-Day Context	Prompt 2 (Mistral) Prompt 5 (Llama) Prompt 3 (Qwen)	Prompt 3 (Mistral) Prompt 5 (Llama) Prompt 4 (Qwen)
Multi-Day Context	Prompt 5 (Mistral) Prompt 1 (Llama) Prompt 2 (Qwen)	Prompt 1 (Mistral) Prompt 2 (Llama) Prompt 3 (Qwen)

Table 5: Highest performing queries on Progress Note generation

Model	Prompt	Method	Setting	ROUGE-L	SapBERT	CUI-F
Mistral	Prompt 4	Multi-Day Context	GEN	25.46	66.47	29.77
Llama3	Prompt 5	Single-Day Context	GEN	18.81	71.47	43.39
Qwen2.5	Prompt 4	Single-Day Context	GT	23.45	71.89	42.26

Table 6: RAG Progress Note highest performing configurations

5.1.2 Progress Note Generation

The overall results for Progress Note generation were significantly higher than Discharge Summary generation, with Qwen2.5-7B and Llama3-8B-Instruct performing best. Their respective results are given in Tables 2 and 3. Methods 1 (Single-Day Context) and 2 (Multi-Day Context) consistently outperform the baseline method, with some models preferring method 1 and others method 2. This indicates that models with higher results using method 2 are likely better at processing very long inputs.

5.2 RAG

Since RAG systems can be very sensitive to query phrasing, query optimization was performed for both tasks, on a scaled-down sample set of admissions. Five prompts were tested for each section - Primary Diagnosis, Hospital Course and Discharge Instructions for the Discharge Summary generation task and Assessment and Plan sections for the Progress Note generation task. The results of this optimization for Discharge Summary generation can be found in Table 4. For Progress Note generation, the highest performing queries are presented in Table 5, and the results of the highest performing configurations per model are given in Table 6. As is the case with the Direct Generation approach, the scores are higher for Progress Note generation, and the values are more consistent here, while the Discharge Summary generation scores are more varied.

Overall, the RAG approach consistently outperforms the Direct Generation approach, especially the Primary Diagnosis CUI f-score and SapBERT BERTScore.

6 Discussion

In this paper, we examined the capabilities of LLMs to summarize clinical text using the tasks of Discharge Summary generation and Progress Note generation. We contrasted the direct generation approach with a RAG one, and determined that a structured RAG-based approach can significantly boost performance on these tasks. Additionally, we examined the effects of different modalities, concluding that free-text note data scores highest, with the inclusion of tabular data actually decreasing performance. Temporal information does not seem to affect performance positively or negatively, with shuffled and chronological data performing similarly. With regard to long context, results from the Progress Note generation task suggests that the models are reasonably capable of handling longer input.

7 Conclusion

Our work examined LLM's capabilities on long-context clinical summarization using a publically available EHR dataset. We found that current LLMs continue to face significant challenges when it comes to accurately summarizing clinical texts, particularly for long-context documents that require nuanced comprehension and coherence. Additionally, these models struggle with temporal reasoning over patient trajectories, making it difficult to track and interpret the sequence of medical events effectively. While RAG has shown some potential in improving performance, the overall results remain substandard, emphasizing the need for further advancements in this area. As part of our future research, we plan to enhance the models' ability to capture and structure critical medical information more effectively.

References

Griffin Adams, Emily Alsentzer, Mert Ketenci, Jason Zucker, and Noémie Elhadad. 2021. What's in a summary? laying the groundwork for advances in hospital-course summarization. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2021, page 4794. NIH Public Access.

Griffin Adams, Jason Zucker, and Noémie Elhadad. 2024. Speer: Sentence-level planning of long clinical summaries via embedded entity retrieval. *arXiv preprint arXiv:2401.02369*.

AI@Meta. 2024. Llama 3 model card.

Kenichiro Ando, Takashi Okumura, Mamoru Komachi, Hiromasa Horiguchi, and Yuji Matsumoto. 2022. Is artificial intelligence capable of generating hospital discharge summaries from inpatient records? *PLOS Digital Health*, 1(12):e0000158.

Emma Croxford, Yanjun Gao, Brian Patterson, Daniel To, Samuel Tesch, Dmitriy Dligach, Anoop Mayampurath, Matthew M Churpek, and Majid Afshar. 2024. Development of a human evaluation framework and correlation with automated metrics for natural language generation of medical diagnoses. *medRxiv*.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, Matthew M Churpek, and Majid Afshar. 2023a. Overview of the problem list summarization (probsum) 2023 shared task on summarizing patients' active diagnoses and problems from electronic health

record progress notes. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2023, page 461. NIH Public Access.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, Samuel Tesch, Ryan Laffin, Matthew M Churpek, and Majid Afshar. 2022a. Hierarchical annotation for building a suite of clinical natural language processing tasks: Progress note understanding. In *LREC... International Conference on Language Resources & Evaluation:[proceedings]. International Conference on Language Resources & Evaluation*, volume 2022, page 5484. NIH Public Access.

Yanjun Gao, Timothy Miller, Dongfang Xu, Dmitriy Dligach, Matthew M Churpek, and Majid Afshar. 2022b. Summarizing patients' problems from hospital progress notes using pre-trained sequence-to-sequence models. In *Proceedings of COLING. International Conference on Computational Linguistics*, volume 2022, page 2979. NIH Public Access.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023b. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2020. Mimic-iv. *PhysioNet. Available online at: https://physionet. org/content/mimiciv/1.0/(accessed August 23, 2021)*, pages 49–55.

Archana Laxmisan, Allison B McCoy, Adam Wright, and Dean F Sittig. 2012. Clinical summarization capabilities of commercially-available and internally-developed electronic health records. *Applied clinical informatics*, 3(01):80–93.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Rui Li, Fenglong Ma, and Jing Gao. 2022. Integrating multimodal electronic health records for diagnosis prediction. In *AMIA Annual Symposium Proceedings*, volume 2021, page 726.

Jennifer Liang, Ching-Huei Tsou, and Ananya Poddar. 2019. A novel system for extractive clinical note summarization using EHR data. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 46–54, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Donald AB Lindberg, Betsy L Humphreys, and Alexa T McCray. 1993. The unified medical language system. *Yearbook of medical informatics*, 2(01):41–51.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2020. Self-alignment pretraining for biomedical entity representations. *arXiv preprint arXiv:2010.11784*.

Mengxian Lyu, Cheng Peng, Daniel Paredes, Ziyi Chen, Aokun Chen, Jiang Bian, and Yonghui Wu. 2024. Uf-hobi at "discharge me!": A hybrid solution for discharge summary generation through prompt-based tuning of gatortrongpt models. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 685–695.

Farida Mohsen, Hazrat Ali, Nady El Hajj, and Zubair Shah. 2022. Artificial intelligence-based methods for fusion of electronic health records and imaging data. *Scientific Reports*, 12(1):17981.

Skatje Myers, Timothy A Miller, Yanjun Gao, Matthew M Churpek, Anoop Mayampurath, Dmitriy Dligach, and Majid Afshar. 2024. Lessons learned on information retrieval in electronic health records: a comparison of embedding models and pooling strategies. *Journal of the American Medical Informatics Association*, page ocae308.

Rimma Pivovarov and Noémie Elhadad. 2015. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, 22(5):938–947.

Joel Shoolin, L Ozeran, Claus Hamann, and W Bria Ii. 2013. Association of medical directors of information systems consensus on inpatient electronic health record documentation. *Applied clinical informatics*, 4(02):293–303.

Christina Silcox, Eyal Zimlichmann, Katie Huber, Neil Rowen, Robert Saunders, Mark McClellan, Charles N Kahn III, Claudia A Salzberg, and David W Bates. 2024. The potential for artificial intelligence to transform healthcare: perspectives from international health leaders. *NPJ Digital Medicine*, 7(1):88.

Robert M Wachter and Erik Brynjolfsson. 2024. Will generative artificial intelligence deliver on its promise in health care? *Jama*, 331(1):65–69.

Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Jason Hom, William J Collins, Ankit Modi, Robert Lloyd, Benjamin Hopkins, et al. 2024. Overview of the first shared task on clinical text generation: Rrg24 and "discharge me!". In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 85–98.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Appendix

What is the patient's main diagnosis?
The patient's primary diagnosis is:
Identify the primary reason for the patient's hospital admission:
Instruct: Given a search query, retrieve relevant passages that answer the query. **Query: patient's primary diagnosis.**
The patient has been diagnosed with:

Table 7: Queries used for retrieving the primary diagnosis. Highest performing queries bolded.

Summarize the hospital course for this patient in a concise and accurate way.
The patient's hospital course included the following:
Provide a brief hospital course, including key events and treatments.
Instruct: Given a search query, retrieve relevant passages that answer the query. Query: Brief hospital course.
What were the key events and outcomes during the patient's hospital stay?

Table 8: Queries used for retrieving the hospital course. Highest performing queries bolded.

Given the input EHR data, generate discharge instructions for this patient.
What are the discharge instructions for the patient?
Write a summary of the discharge plan, including medications, follow-up visits, and patient care instructions.
Instruct: Given a search query, retrieve relevant passages that answer the query. Query: discharge instructions.
What follow-up care and medications are recommended for the patient after discharge?

Table 9: Queries used for retrieving the discharge instructions. Highest performing queries bolded.