Roux-lette at “Discharge Me!”: Reducing EHR Chart Burden with a Simple, Scalable, Clinician-Driven AI Approach

S. Wendelken[*], A. Antony[**], R. Korutla[*], B. Pachipala[*], D. Mahajan[**], J. G. Shanahan[**], W. Saba[**]
[*]The Roux Institute at Northeastern University
[**]The Institute for Experiential AI at Northeastern University

Abstract

Healthcare providers spend a significant amount of time reading and synthesizing electronic health records (EHRs), negatively impacting patient outcomes and causing provider burnout. Traditional supervised machine learning approaches using large language models (LLMs) to summarize clinical text have struggled due to hallucinations and lack of relevant training data. Here, we present a novel, simplified solution for the “Discharge Me!” shared task. Our solution uses a question-based approach to treat this summarization task as a context-aware and domain-specific question-answering process. Our pipeline prompts an LLM answer *specific questions* posed by subject-matter experts (SMEs) using only patient specific context data. This method (i) avoids hallucinations through hybrid RAG/zero-shot contextualized prompting; (ii) requires no extensive training or fine-tuning; and (iii) is adaptable to various clinical tasks.

1 Introduction

Clinicians spend considerable amount of time navigating vast amounts of electronic health records (EHRs) information, often spending 2-3 times more time interacting with EHRs than with patients [1][4]. Much of this time is spent on manual information retrieval and stylized summarization of relevant content, referred to as the charting burden. The "Discharge Me!" shared NLP task [8][9] aims to alleviate this burden by automating the generation of "Brief Hospital Course" (BHC) and discharge instructions from EHRs in the MIMIC IV dataset [8]. These sections are critical but time-consuming, requiring synthesis, interpretation, and summarization of data from various parts of the patient’s medical charts include admission notes, progress notes, lab results, and radiology reports.

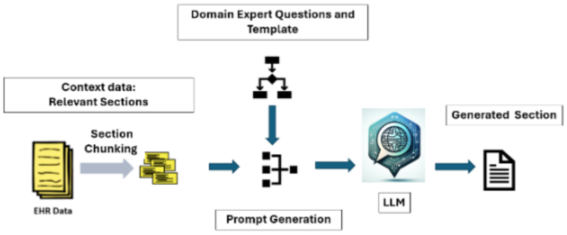

Domain Expert Questions and Template
Context data: Relevant Sections
Section Chunking
EHR Data
Prompt Generation
LLM
Generated Section

***Figure 1**: Simplified solution to “Discharge Me!” task.*

Automated approaches such as rule-based systems and supervised machine learning models have been explored but often struggle with the complexity and variability of clinical data. These challenges lead to issues like hallucinations—plausible but incorrect information—and incomplete summaries due to missing data in the training sets. Despite significant progress in machine learning for NLP tasks, achieving contextualized, relevant, and accurate summarization remains challenging. Supervised ML approaches face difficulties in domain adaptation, and fine-tuning does not fully address these issues [1][3][5]. Moreover, and while some progress was reported using fine tuning (e.g., [11]), the approach we describe here achieves better results without bearing the enormous cost of fine tuning. The lack of training data and LLMs' limitations in handling domain-specific data further complicate the problem. Additionally, different users require different types of information extracted from EHRs, making contextual summarization essential.

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 719–723
August 16, 2024. ©2024 Association for Computational Linguistics

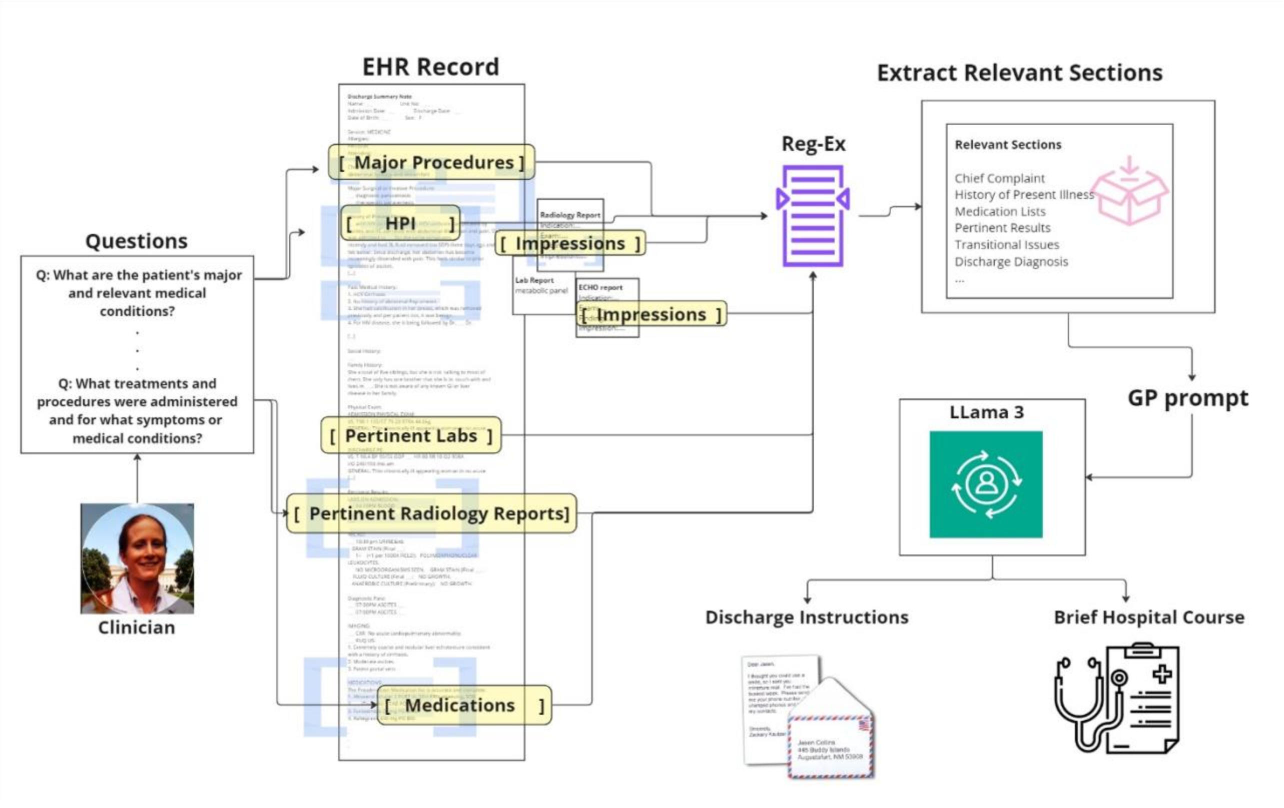


Figure 2: *"GP prompting" pipeline splits the EHR record into substring contexts to which domain expert questions are applied in the form of prompting. The responses to these questions are then used to generate a task centric response.*

Our solution uses a question-based approach to treat this summarization task as a context-aware and domain-specific question-answering process (see Figure 1). The summaries generated using our pipeline answer *specific questions* posed by subject-matter experts (SMEs). Previous experiments of our approach for lengthy clinical note summarization using general-purpose models like ChatGPT showed their ability to understand clinical terms and abbreviations without fine-tuning [6]. This supports our hypothesis that LLMs can efficiently generate accurate summaries without hallucinations when provided with the relevant contextual data. Our method is customizable, contextual, and bypasses the need for extensive training datasets. It harmonizes unstructured clinical data and can be applied to any domain requiring contextualized summaries for various users.

2 Methods

We developed an AI pipeline to automate clinical workflow tasks for generating discharge summaries, leveraging large language models (LLMs) to answer questions, retrieve information, infer actions, and summarize the information in a stylized format, where a domain expert guided the questions and output style (See Figure 2). We call our approach "GP-Prompting". Our first iteration used the general-purpose LLM Meta-Llama-3-8B-Instruct [10] installed locally on a A10G GPU.

The following is a summary of our pipeline:

1. *Data Sources*: A subset of the MIMIC IV data set including Discharge Summary text notes, radiology reports, and initial ICD diagnosis codes provided by the task organizers [8].
2. *Pre-processing*: Segmentation of discharge note into logical sections. This was done by extracting content by section, where section headings were identified from the dataset by the clinician. Section headers were not standardized in the data set and were thus matched to known variations using regexp. E.g., "History of Present Illness" | "HPI". Extraneous text (e.g., repeated "=") were removed
3. *"GP-prompting"*: Our method is a hybrid of question answering / zero-shot prompting.
   a. Domain Expert Input: questions formulated by clinician to identify key medical information.
   b. Select relevant sections of the clinical notes data that will contain answers to the

questions suggested by a domain expert. E.g. Select the "HPI" section to answer questions about the initial treatment course. Use the "Admission and Discharge Medications" lists to deduce what new medications the patient was prescribed for discharge. The selected text will be used as the context data.

c. Prompt Template: Prompts were constructed by concatenating a {Topic}, {General Instructions}, {Output Template}, {SME Questions}, {Text Generation Instructions}, {Context} data.

Topic: "The topic is about clinical notes, medical records, and other text documents from electronic health records (EHR) from a patient's hospital admission."

General Instructions: "You will be provided with input text and data from a specific patient's medical records. Use only this data to answer questions about the EHR.

"Discharge Instruction" generation

Output Template:

": "Dear ____, It was a pleasure to take care of you during your recent hospital admission. You were admitted to the hospital because [explain the reason for admission]. [Briefly explain the diagnostic work up and explain the results] During your hospital stay your treatments included [briefly explain the major treatments and procedures]. … Sincerely, Your Team"

SME Questions:

"Why was the patient admitted to the hospital?
What treatments and procedures were administered and for what symptoms or medical conditions?
What was the diagnostic work up related to the chief complaint? What were the notable results?
What medications were used to treat the patient? Were any medications new or discontinued?
Are there changes to the existing medications?
What are the ongoing issues and follow-up recommendations?"

Text Generation Instructions: ": Write a letter to the patient that summarizes their hospital stay and communicates follow-up instructions as well as important changes to their medications. Use the provided template and answers to the questions above to fill in the blanks.

Selected Context: 'CC', 'HPI', 'Transitional Issues', 'Discharge Diagnosis' from the discharge summary note.

Example "Discharge Instructions" output:

"Discharge Instructions ---hadm_id: 27729294

Dear ___, It was a pleasure to take care of you during your recent hospital admission. You were admitted to the hospital because of your leg swelling and complaints of oral ulcers. During your hospital stay, your treatments included antibiotics for cellulitis in your leg and a medication called Duonebs. You were diagnosed with cellulitis in your leg and were given a course of antibiotics to help clear up the infection. You were also treated for your bullous pemphigoid and lymphoma in your mouth. You will need to follow up with your doctor to monitor your condition and make sure that the infection has cleared up. You will also need to continue taking your medication as prescribed and follow up with your dermatologist to manage your bullous pemphigoid. If your symptoms return, please call our office (555-555-5555) or 911 for immediate assistance. Sincerely, Your Team---"

Brief Hospital Course (BHC) Generation:

Output Template: "___ with a history of [briefly describe the relevant medical history or existing medical problems] presents with [briefly summarize the patient's symptoms]. They were admitted to the [what unit, service or floor were they admitted to?] for [briefly summarize the reason for admission, initial diagnosis, and any major procedures or treatments]. [Briefly summarize the discharge diagnosis at discharge]

Problem-based hospital course—

# [Medical condition 1] --- [briefly describe presentation and severity for this medical condition or symptom]. [Summarize the related diagnostic work up and notable results]. [Summarize the related medical or surgical treatment course].

# [Medical condition 2] -- [briefly describe presentation and severity]..."

SME Questions for BHC:

"What are the patient's major medical conditions? Why was the patient admitted to the hospital? What symptoms did the patient present with? What was the severity and duration of the presenting symptoms? What treatments and procedures were administered and for what symptoms or medical conditions? What unit or floor were they admitted to? What was the diagnostic work up? What were the notable results? What medications were used to treat the patient? Were any medications new or discontinued? Are there changes to the existing medications? What was the patient diagnosed with at discharge? What was the patient initially diagnosed with?"

Text Generation Instructions: “Fill in the blanks: Use the answers to the questions above to fill in the blanks. If you do not know the answer, fill in the blank with "___". Answer clearly and concisely”
Context: text data included “Chief Complaint”, “HPI”, “Discharge Diagnosis,” and “Medication Lists” from the discharge note, and “ICD Diagnosis title” from the diagnosis table
Example BHC output for hadm_id 27729294: “The patient with a history of bullous pemphigoid and lymphoma presents with leg swelling and complaints of oral ulcers. They were admitted to the medical floor for evaluation of leg lesions and abdominal swelling, initially diagnosed with cellulitis of the left lower limb.

Problem-based hospital course---

# Cellulitis --- The patient presented with leg swelling and complaints of oral ulcers. The diagnostic work up included labs notable for WBC of 12.5, and imaging such as LLE US did not show evidence of DVT; CXR showed mild cardiomegaly and mild vascular congestion. The patient was given Duonebs and CTX 1g IV for LLE cellulitis.
# Bullous pemphigoid --- The patient has a history of bullous pemphigoid and lymphoma in his mouth. The diagnostic work up included bedside US did not show any evidence of ascites.”

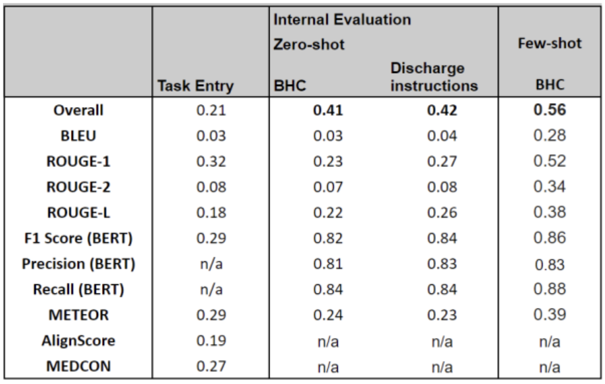
3 Evaluation

Outputs for all 10,962 targets were submitted to the shared task and evaluated against a hidden subset of 250 records by the task organizers as described in [9]. Results are presented in Table 1. Concurrently, we evaluated the first 250 generated outputs against human-generated targets using BLEU-4, ROUGE-1/-2/-L, BERTScore (Precision, Recall, F1), and METEOR scores. A clinician visually inspected the first 10 outputs for accuracy and recall during pipeline development.

Time and computational constraints limited our ability to fully optimize the pipeline during the contest. Post-contest, we tested a few-shot approach for a single record with five example prompt-data pairs, showing promising results for BHC generation.

4 Results

Initial results showed promising performance, though further optimization with few-shot learning and refined model parameters could improve accuracy and efficiency.



		Internal Evaluation		
		Zero-shot		Few-shot
	Task Entry	BHC	Discharge instructions	BHC
Overall	0.21	**0.41**	**0.42**	**0.56**
BLEU	0.03	0.03	0.04	0.28
ROUGE-1	0.32	0.23	0.27	0.52
ROUGE-2	0.08	0.07	0.08	0.34
ROUGE-L	0.18	0.22	0.26	0.38
F1 Score (BERT)	0.29	0.82	0.84	0.86
Precision (BERT)	n/a	0.81	0.83	0.83
Recall (BERT)	n/a	0.84	0.84	0.88
METEOR	0.29	0.24	0.23	0.39
AlignScore	0.19	n/a	n/a	n/a
MEDCON	0.27	n/a	n/a	n/a

Table 1: *Scoring metrics for “Discharge Me!” generated outputs “Brief Hospital Course” and “Discharge Instructions”*

5 Discussion

Our approach demonstrated a simple and effective method for automatically generating the “Brief Hospital Course” and “Discharge Instructions” sections of discharge summary notes. Future improvements include integrating few-shot learning, fine-tuning, and principled chunking with retrieval-augmented generation (RAG). Experimentation with various LLM sizes and optimization of parameters (e.g., temperature, different values for *top_k*), topic tracking, and integration of structured chart data (not available for this task) can enhance output quality and speed.

6 Limitations

The task dataset was unrealistic, lacking essential components present of typical charts such as daily progress notes, procedure notes, labs, vitals, microbiology, radiology reports, and medication administration records (MAR). Generating the BHC and discharge instructions without comprehensive event data leads to hallucinations. The provided dataset, containing only discharge summary notes, is insufficient for accurate BHC or discharge instructions, especially for patients with extended hospital stays.

Additionally, the target dataset sections were often inaccurately segmented from the input. Approximately 16% of phase 2 BHC targets were severely incomplete, often under 100 words. In these cases, the extraction was truncated due to an unexpected heading, often missing the problem-based treatment course entirely. The targets also often incorrectly included content from the "Transitional Issues" section, which should be separate from the BHC.

We lacked comprehensive data such as daily progress notes and outpatient referrals, so we utilized selected parts of the discharge summary, including the HPI Medication list, which provided partial relevant information needed for the BHC. All selected input sections were considered by the clinician to be accessible during the typical clinical workflow. Incomplete records often resulted in outputs lacking the full content of the target data, but it was reassuring that the model did not hallucinate.

Pipeline Challenges: Due to data-use agreements, models and data had to be run locally and securely, necessitating downloaded LLMs. This limitation prevented the use of faster, publicly available pipelines, resulting in lower accuracy, and slower local model outputs compared to more advanced models that we plan to use in the future.

We also noted a discrepancy between contest and internal BERTScores. At the time of this publication, the root cause of this discrepancy is unknown, but it is likely resulting from using different BERTScore functions (we used a standard "bert_score" import, whereas the contest scoring used a custom BERTScore script). Similarly, the custom AlignScore and MEDCON scores used for contest were not implemented during our evaluation process as we were unable to successfully run the custom scripts in time for the contest entry.

7 Conclusion

The solution we presented was an efficient, context-aware, question-based approach to automate the generation of discharge summaries. Despite the constraints and limitations of the dataset and evaluation metrics, our method showed promise, particularly with a few-shot learning approach. Future work will focus on refining chunking methods for a RAG-based approach, optimizing prompts, and exploring various LLM configurations to improve accuracy and reliability in clinical settings.

References

[1] E. Hossain et al., "Natural Language Processing in Electronic Health Records in relation to healthcare decision-making: A systematic review," Comput. Biol Med., vol. 155, p. 106649, Mar. 2023, doi: 10.1016/j.compbiomed.2023.106649.

[2] T. Searle, Z. Ibrahim, J. Teo, and R. J. B. Dobson, "Discharge summary hospital course summarization of in-patient Electronic Health Record text with clinical concept guided deep pre-trained Transformer models," J. Biomed. Inform., vol. 141, p. 104358, May 2023, doi: 10.1016/j.jbi.2023.104358.

[3] D. Van Veen et al., "Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts," Res. Sq., p. rs.3.rs-3483777, Oct. 2023, doi: 10.21203/rs.3.rs-3483777/v1.

[4] J. M. Overhage and D. McCallie, "Physician Time Spent Using the Electronic Health Record During Outpatient Encounters," Ann. Intern. Med., vol. 172, no. 3, pp. 169–174, Feb. 2020, doi: 10.7326/M18-3684.

[5] T. Searle, Z. Ibrahim, and R. J. Dobson, "Experimental Evaluation and Development of a Silver-Standard for the MIMIC-III Clinical Coding Dataset," in Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing, 2020, pp. 76–85. doi: 10.18653/v1/2020.bionlp-1.8.

[6] W. Saba, S. Wendelken, and J. Shanahan, "Question-Answering Based Summarization of Electronic Health Records using Retrieval Augmented Generation." Preprint https://arxiv.org/ftp/arxiv/papers/2401/2401.01469.pdf

[7] A. S. Afshar et al., "An exploratory data quality analysis of time series physiologic signals using a large-scale intensive care unit database," JAMIA Open, vol. 4, no. 3, p. ooab057, Jul. 2021, doi: 10.1093/jamiaopen/ooab057.

[8] J. Xu, "Discharge Me: BioNLP ACL'24 Shared Task on Streamlining Discharge Documentation." [object Object]. doi: 10.13026/4A0K-4360. https://physionet.org/content/discharge-me/1.2/

[9] J. Xu et al., "Overview of the First Shared Task on Clinical Text Generation: RRG24 and 'Discharge Me!,'" in The 23rd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024.

[10] AI@Meta, "meta-llama/Meta-Llama-3-8B-Instruct · Hugging Face." Accessed: May 16, 2024. [Online]. Available: https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct.

[11] Pal K, Bahrainian SA, Mercurio L, Eickhoff C, 2023, "Neural Summarization of Electronic Health Records", JMIR Preprints, DOI 10.2196/preprints.49544.