Utilizing a RAG-LLM as a Physician's Assistant

Edgar Leon, Steven Lu, Monica Mathur,
Omkar Salpekar, Emily Zhou

Berkeley
UNIVERSITY OF CALIFORNIA

Introduction

- **Problem:** Physicians spend 60-66%* of their time sifting through medical records/notes to understand and summarize key information, rather than focusing on their core competency (providing care)
- **Impact of solving this problem:**
  - Alleviates the documentation burden that delays determining diagnoses, and formulating treatment plans
  - Has the potential to impact over one million U.S. physicians
- **Solution:** A Retrieval Augmented Generation (RAG) - LLM physician's assistant that enables the retrieval of specific, accurate, and relevant information

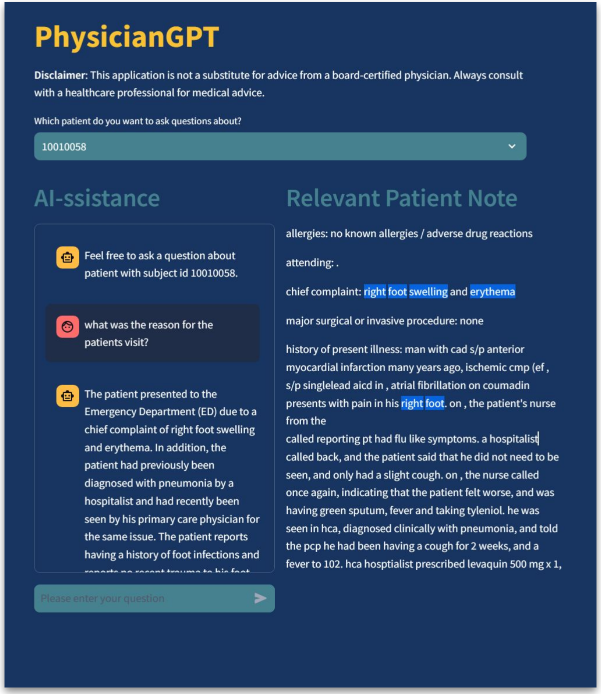
Documentation burden 60-66% of their time

Impacts over one million physicians in the US

Berkeley
UNIVERSITY OF CALIFORNIA

*https://arxiv.org/pdf/2309.07430.pdf

MVP Demo



PhysicianGPT

Disclaimer: This application is not a substitute for advice from a board-certified physician. Always consult with a healthcare professional for medical advice.

Which patient do you want to ask questions about?

10010058

AI-ssistance

Feel free to ask a question about patient with subject id 10010058.

what was the reason for the patients visit?

The patient presented to the Emergency Department (ED) due to a chief complaint of right foot swelling and erythema. In addition, the patient had previously been diagnosed with pneumonia by a hospitalist and had recently been seen by his primary care physician for the same issue. The patient reports having a history of foot infections and

Relevant Patient Note

allergies: no known allergies / adverse drug reactions

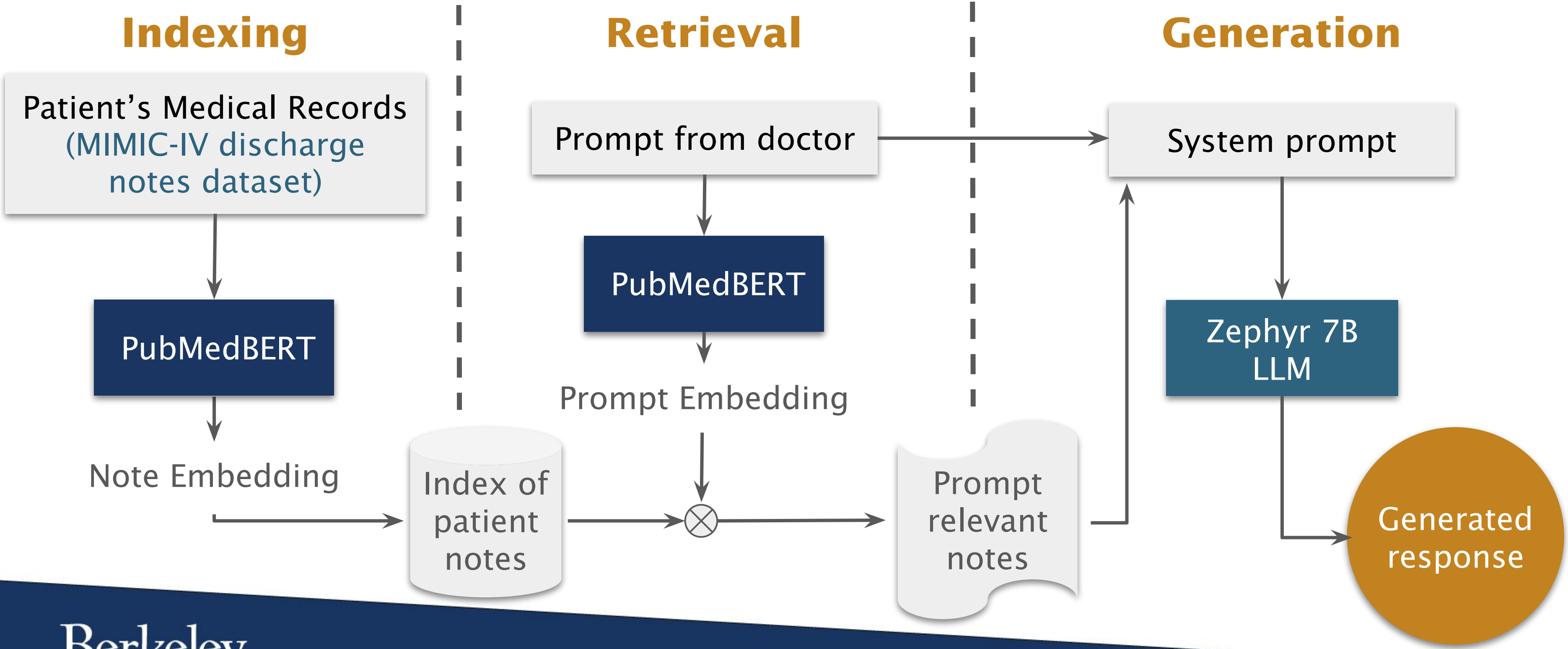
attending: .

chief complaint: right foot swelling and erythema

major surgical or invasive procedure: none

history of present illness: man with cad s/p anterior myocardial infarction many years ago, ischemic cmp (ef , s/p singlelead aicd in , atrial fibrillation on coumadin presents with pain in his right foot. on , the patient's nurse from the
called reporting pt had flu like symptoms. a hospitalist called back, and the patient said that he did not need to be seen, and only had a slight cough. on , the nurse called once again, indicating that the patient felt worse, and was having green sputum, fever and taking tyleniol. he was seen in hca, diagnosed clinically with pneumonia, and told the pcp he had been having a cough for 2 weeks, and a fever to 102. hca hosptialist prescribed levaquin 500 mg x 1,

Retrieval-Augmented Generation (RAG)



Berkeley
UNIVERSITY OF CALIFORNIA

The Chunking Strategy

INPUT

note_text: string

"Discharge Condition: Mental Status: Clear and coherent. Level of Consciousness: Alert and interactive. Activity Status: Ambulatory - Independent."

CHUNK STRATEGY

(example)
Chunk size = 5
Overlap = 2

(actual)
Chunk size = 1024
Overlap = 100

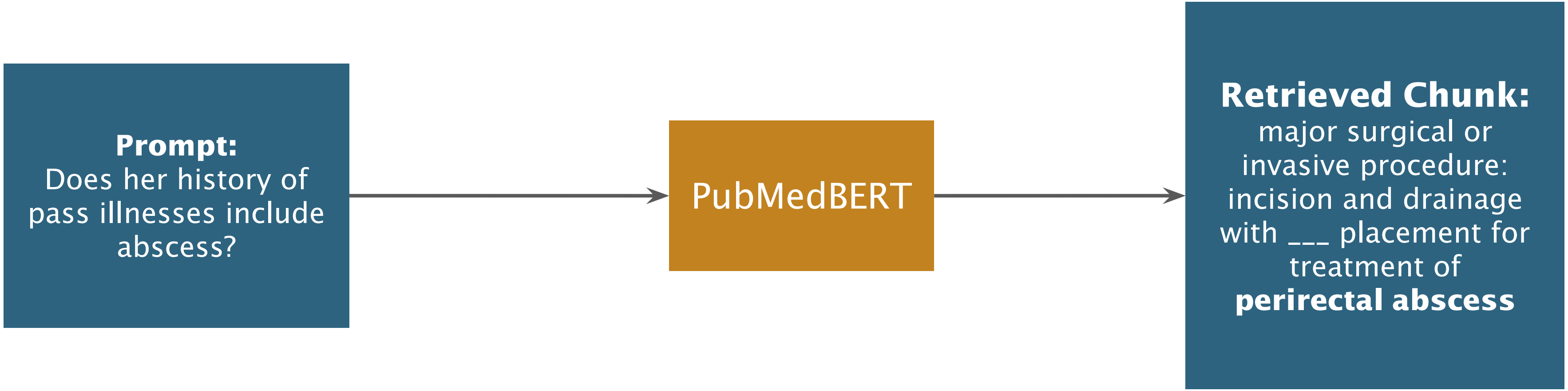
OUTPUT

chunks: list[str]

["Discharge Condition: Mental Status: Clear", "Status: Clear and coherent. Level", … , "interactive. Activity Status: Ambulatory - Independent."]

Retrieval: Selection of LLM

PubMedBERT (**is** trained on medical content)



Prompt:
Does her history of pass illnesses include abscess?

PubMedBERT

Retrieved Chunk:
major surgical or invasive procedure: incision and drainage with ___ placement for treatment of **perirectal abscess**

Key learning: combination of **domain** expertise and **task training** optimization **drastically improves** retriever **performance**.

Generator: LLM & System Prompt Selection

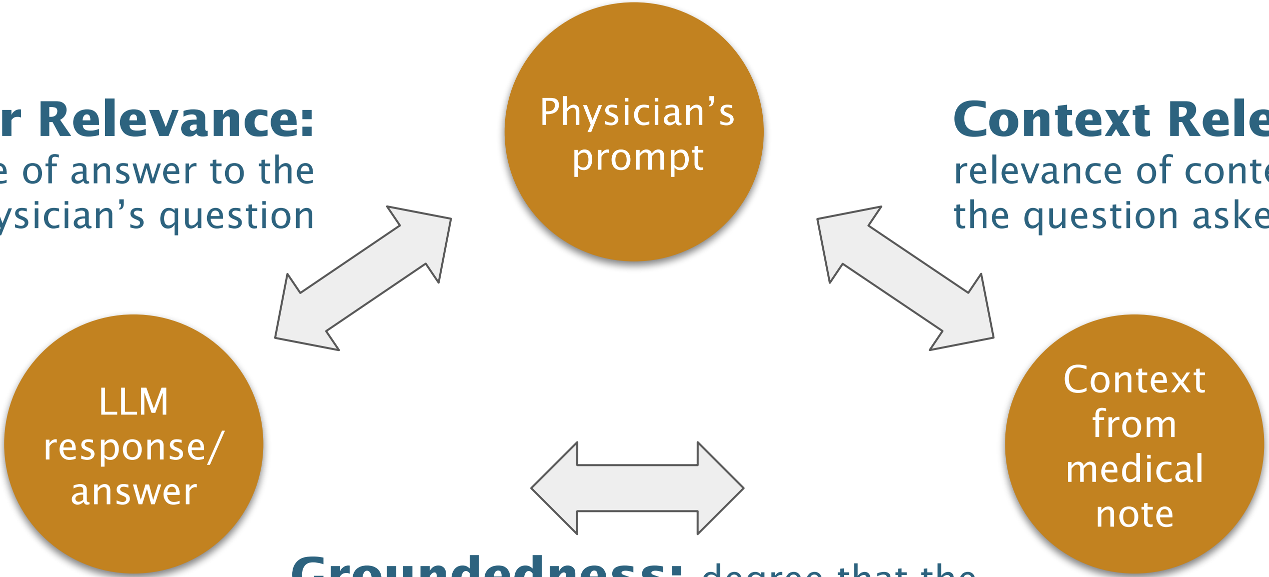
Generator LLM

- LLM: Zephyr-7B (fine-tune of Mistal 7B)
- Key Learning: Choose the best aligned chat model (RLHF/DPO/etc.)

System Prompt Engineering

- Initial System Prompt: *“Please answer {question} given the following context: {note}”*
- Final System Prompt: *“You are an expert doctor. I am giving you the following excerpt from a patient’s medical record: {note}. Please use only the excerpt to clearly, concisely, and confidently answer the question.”*
- Key Learning: Instill confidence, provide context, discourage use of non-contextual info

Evaluation Framework



Physician's prompt

LLM response/ answer

Context from medical note

Answer Relevance: relevance of answer to the physician's question

Context Relevance: relevance of context chunk to the question asked

Groundedness: degree that the answer provided by the LLM is supported by the context from the medical note

**Underlying principles of the evaluation framework is similar to the TruLens RAG triad*
https://www.trulens.org/trulens_eval/getting_started/core_concepts/rag_triad/

Test Evaluation Dataset

Method of Generation:

- 30 observations generated manually by reviewing the MIMIC-IV notes dataset

Attributes:

- Patient ID
- Note ID
- Question
- Ground Truth Chunk
- Ground Truth Answer

Question Make-up:

- Categories: Family history, Medications, Complaints & Diagnosis
- Type: Yes/No vs. open-ended

Evaluation Examples - Retrieval Metrics

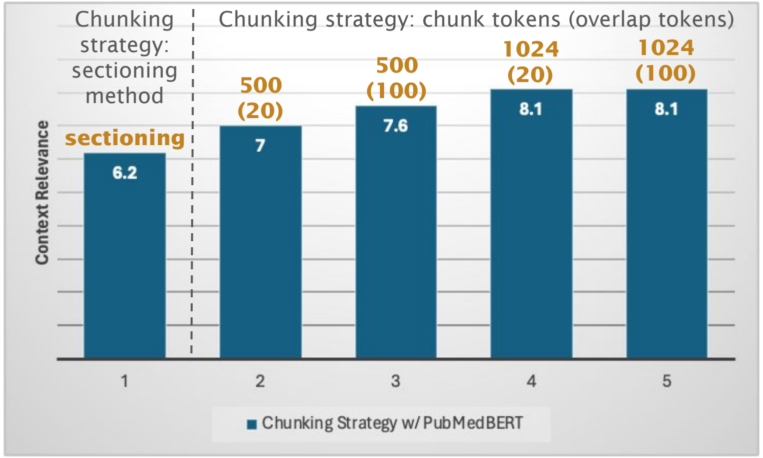
Question	Chunk method	Chunk retrieved	Context Relevance
Was there any hemorrhaging observed in the head CT scan?	Sectioning method	Family History Section	0
	1024 tokens with 100 tokens overlap	Most of the medical note, i.e. multiple sections	10

Evaluation Examples - Generation Metrics

Question	Ground truth answer	Answer provided	Score*
What medications were used to treat her heart condition?	Two medications were used to treat the patient’s heart condition: atorvastatin 80 mg daily and metoprolol 25 mg daily.	The new medications prescribed to treat her heart condition were atorvastatin 80 mg daily and metoprolol 25 mg daily.	9.1 (8.8)
Does the patient have any allergies?	Yes, the patient has optiflux dialyzer as an allergy.	Yes, the patient has an allergy to optiflux dialyzer, as mentioned in the excerpt provided.	8.1 (9.8)
What was the patient's discharge condition?	The patient was discharged with a mental status that was clear and coherent, with a consciousness that was alert and interactive, and an ambulatory activity status of independent.	I do not have access to the full medical record or the current updated condition of the patient. however, based on the excerpt provided, it does not mention a discharge date or the patient's discharge condition. kindly refer to the patient's comprehensive medical record or contact the healthcare facility where the patient was last treated to get more information about the patient's discharge condition.	0.56 (0.67)

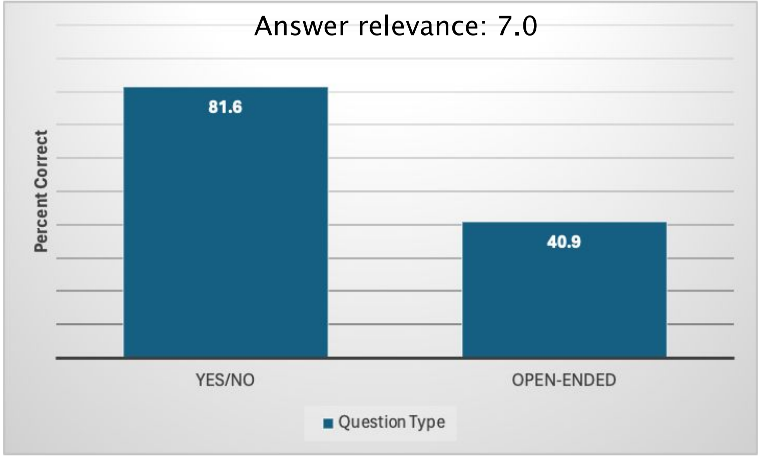
RAG System Performance

Effect of Chunking Strategy on Context Relevance



Key learning: Best performance achieved with large chunks, vs. sectioning method → a large chunk ensures context is present

Effect of Question Type on Generator Performance*



Key learning: Best performance achieved with yes/no questions; Incomplete, wordy responses, or retrieved chunk pulled from the wrong note are areas of improvement

**all with PubMedBERT, and 1024 tokens/100 overlap*

Berkeley
UNIVERSITY OF CALIFORNIA

Scores based on an average of the 30 observations included in the test dataset

Feedback from Board Certified Physicians

Product Strengths:

- Answers vast majority of questions correctly and accurately
- Provides helpful context from history

Capability Enhancement Opportunities:

- Address sensitivity to word choices
    - Medications “on admission”
    - “Does the patient smoke?” vs. “Does the patient have a history of smoking?”
- Improve understanding of chronology of events
- Add capability to retrieve context from multiple notes

Our mission is to …

Treat patients with care, faster!