

# MediMate: Medical Q&A App

Team 1: Hyrum Eddington, Sanidhya Singal, Shyam Renjith

# Outline

**1** Introduction and Motivation

**2** Datasets

**3** Methodology

**4** Experiments and Results

**5** Demonstration

**6** Conclusion

# Introduction and Motivation

# Chatbots in Medical Contexts

- Chatbots have the potential to democratize aspects of the patient experience:
  - Can function unsupervised without clinician supervision
  - Are available at times and in areas where a physician cannot be
  - Easily accessible through computer, smartphone, or other electronic device
- Accurate, effective chatbots for these same reasons can also reduce physician task load, such as spending large amounts of time responding to patient queries
- Information can be considered as a social determinant health, meaning tools that increase health literacy can ameliorate the gap in health outcomes among disadvantaged groups

# LLMs and Hallucination

- While LLMs are highly sophisticated in mimicking human speech and text, they can produce inaccuracies or worse, produce misinformation called ‘hallucinations’
- For chatbots that are going to serve in high-stake contexts such as medical consultation, the probability of hallucination should be reduced as much possible to ensure accurate medical information is given

## Brief Report

April 27, 2023

### Performance of an Artificial Intelligence Chatbot in Ophthalmic Knowledge Assessment

Andrew Mihalache, BMSc(C)<sup>1</sup>; Marko M. Popovic, MD, MPH(C)<sup>2</sup>; Rajeev H. Muni, MD, MSc<sup>2,3</sup>

**Findings** In this cross-sectional study that included 125 text-based multiple-choice questions provided by the OphthoQuestions free trial for board certification examination preparation, ChatGPT answered 58 questions (46%) correctly.

# Motivation - Combining LLMs with RAG for Accurate Chatbot Responses

- Retrieval-augmented generation (RAG) extends the capabilities of LLMs by retrieving information related to a patient's query from a large database of medical information and providing that additional context to the LLM
- Advantages include:
  - Can act as a 'guardrail' for the LLM, reducing the risk that hallucinations occur
  - Adds additional, more specific information regarding the medical query based on rich databases compiled using medical research
  - Can be constantly improved as additional data is added—potential to keep 'up-to-date' with medical standard of care, much more efficient than re-training an LLM

# Datasets

# Sources for RAG

## PubMedQA

- Dataset composed of 62K QA instances based on PubMed papers including results and conclusions that can be extracted and used as context for RAG
- Adds rich and precise scientific information to can improve LLM response

## NFCorpus

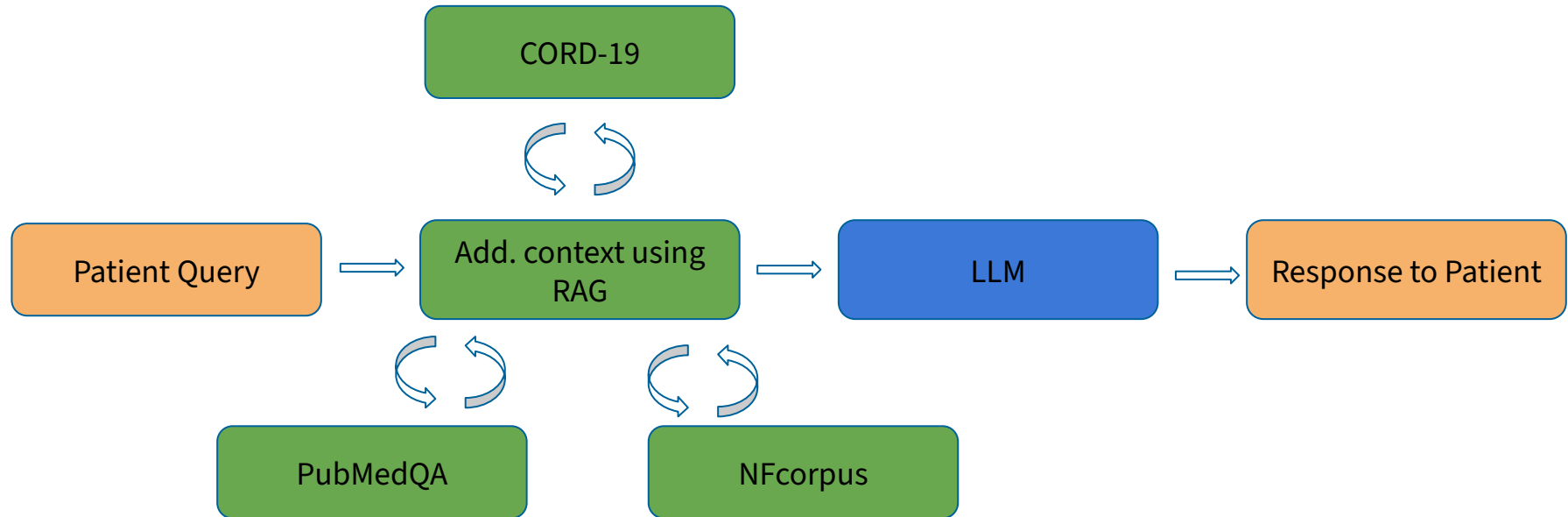
- Dataset based on ~10,000 medical documents including PubMed papers related to general health, and nutrition
- Potential to aid patients and consumers seeking guidance on disease or dietary Q's potentially relevant to their conditions

## CORD-19

- Thousands of COVID-19 related papers up-to-date through June 2022
- Example of using recent research to 'update' context for an LLM–Chatgpt's cutoff year is 2021
- Can provide actual retrieved results from COVID-relevant studies



# From Query to RAG/LLM to Response



# Methodology

# Background

# Deep Learning Embeddings / Vectors

- Numerical representations of data in a lower-dimensional space, learned through Artificial Deep Neural Networks.
- Can represent words, sentences, or documents in a continuous vector space.
- Traditional methods, like one-hot encoding, lack semantic information and struggle with capturing relationships between words.
- Deep learning embeddings are dense vector representations that aim to capture semantic similarities between words.
- Examples:
  - OpenAI Embeddings: text-embedding-ada-002-v2 model
  - LLAMA Embeddings: llama-2-7b-chat model

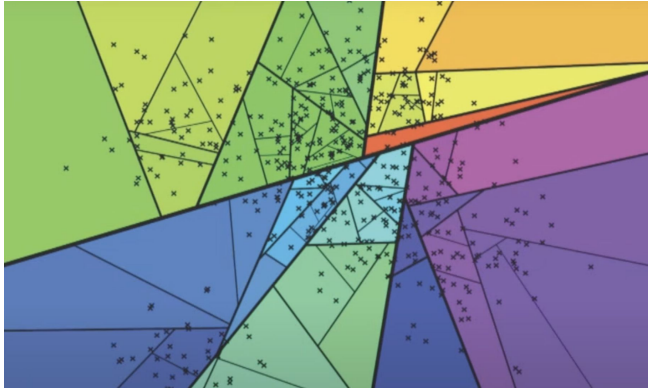
# Vector DBMSs

- Numerical vectors are computationally inexpensive to store and compare.
- Data has high dimensionality and sparsity.
- Vector DBMSs provide efficient storage, retrieval, and processing of vector data.
- VDBMSs support similarity search through indexing methods that enable rapid and accurate searching of similar vectors based on specific distance metrics such as Euclidean distance or cosine similarity.
- Examples:
  - ANNOY: Approximate Nearest Neighbors Oh Yeah
  - FAISS: Facebook AI Similarity Search

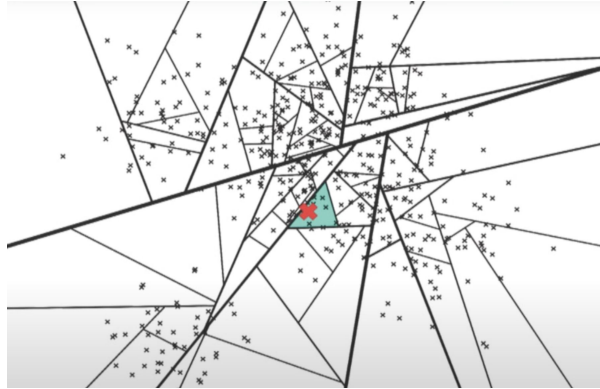
# Approximate Nearest Neighbors

- Problem: Given a query, retrieve the top-k relevant documents.
- Classical Approach:
  - Compute the distance between every document in the dataset and the given query.
  - Computationally expensive to searching for the exact NNs.
- ANN Approach:
  - Reduce the candidate document search space by introducing randomness.
  - ANNs methods trade off accuracy for gaining computational efficiency.

# ANN Example: Annoy



Random division of Space



Documents in Query's Subspace



Reduced Set of Candidate Documents

# **Our Approach**



# Retrieval Augmented Generation

- RAG is a SOTA method for Closed Domain Question Answering. It combines an information retrieval component with an LLM for text generation.
- RAG takes an input and retrieves a set of relevant/supporting documents given a source . The documents are concatenated as context with the original input prompt and fed to the text generator which produces the final output. This helps reduce the problem of “hallucinations” in LLMs.

```
chain = ({ "context": retriever | format_docs, "question": RunnablePassthrough() }  
        | prompt  
        | llm  
        | StrOutputParser())
```

# Our Approach

## 1. LLM without RAG

- Directly pass the user query to the LLM
- “Open Domain Question Answering”

## 2. LLM with RAG

- Find top-k relevant documents from the selected dataset based on user query
- Concatenate the relevant documents and pass them as additional context to the LLM, along with the user query
- “Closed Domain Question Answering”

# User Interface (Demo Later)

- Simple and seamless UI to facilitate typical patient-practitioner conversations.
- Observe response difference in both with and without RAG cases on multiple models and domain-specific datasets.
- Multi-modal input (text, file, speech)
- Serves as a POC for further specialized scalable Medical Q&A systems.

# UI/UX Features

1. Generating Response With and Without RAG
2. Recognizing and Transcribing Speech
3. Switching Across Datasets
4. Maintaining Conversation History
5. Uploading Medical History as a File

# Experiments and Results

# Implementation Details

- OpenAI Embeddings: OpenAI's API based on text-embedding-ada-002 model
- Retriever: FAISS Vector DBMS using `langchain.vectorstores`
  - Unit vectors with dot product metric
  - Retrieve top-k documents where  $k=5$
- LLM: OpenAI LLM using `langchain.llms` based on:
  - GPT-3.5-turbo-instruct model
  - Davinci-0.0.3 model (legacy baseline)
- UI: `React.js`
- Backend: `Flask`

# Metrics

- **BLEU Score (Bilingual Evaluation Understudy):**

- Evaluates the quality of machine-generated text by comparing it to one or more reference texts.
- Involves counting the number of overlapping n-grams (typically up to 4-grams) between the generated and reference texts.

- **ROUGE Score (Recall-Oriented Understudy for Gisting Evaluation):**

- Evaluates the quality of summaries by comparing them to one or more reference summaries.
- Provides a comprehensive measure of the overlap between system-generated summaries and human reference summaries.

$$\text{BLEU} = \min \left( 1, \frac{\text{output-length}}{\text{reference-length}} \right) \left( \prod_{i=1}^4 \text{precision}_i \right)^{\frac{1}{4}}$$

$$\text{ROUGE-N} = \frac{\sum_{S \in \text{reference-summaries}} \sum_{n\text{-gram} \in S} \text{count}_{\text{match}}(n\text{-gram})}{\sum_{S \in \text{reference-summaries}} \sum_{n\text{-gram} \in S} \text{count}(n\text{-gram})}$$

# Results

Results without RAG:

Dataset	Mean BLEU Score	Mean ROUGE Score
NFCorpus	0.7441	0.2167
PubMedQA	0.2721	0.2189
CORD19	0.4709	0.1713

Results with RAG:

Dataset	Mean BLEU Score	Mean ROUGE Score
NFCorpus	0.8377	0.2226
PubMedQA	0.6937	0.2739
CORD19	0.6099	0.2024

- All experiments were conducted with GPT-3.5-turbo-instruct as the LLM.
- For RAG, we used FAISS as the VectorDB for IR.



# Inferences

- RAG consistently boosts performance across all datasets (NFCorpus, PubMedQA, CORD-19).
- Combining RAG with LLM shows notable enhancements in BLEU and ROUGE scores for all datasets, highlighting the importance of this approach.

# Demonstration



## MediMate

Reset Chat
OpenAI gpt-3.5-turbo
☒ Enable RAG
PubMedQA

Hello! I'm MediMate, your biomedical chatbot. I provide information on health topics, assist with medical inquiries, and support healthcare needs. Need insights on medical conditions or health advice? I'm here to help. How can I assist you today?

Do nomograms designed to predict biochemical recurrence (BCR) do a better job of predicting more clinically relevant prostate cancer outcomes than BCR?

Based on the additional context provided, it seems that nomograms designed to predict BCR may not be as accurate in predicting more definitive endpoints such as prostate cancer-specific mortality (PCSM). Further research and analysis would be needed to determine the effectiveness of these nomograms in predicting PCSM.

# Conclusion

# Conclusion

- Formulated the task as a CDQA problem. Researched relevant and suitable medical data for IR and subsequently RAG.
- Implemented RAG for improved accuracy and contextual understanding in domain-specific Q&A using language models and external data.
- Developed intuitive, accessible, and multi-modal UI to maintain seamless conversation-like Q&A sessions.
- Leveraged advanced language models and information retrieval for accurate, relevant responses to complex domain-specific queries.

# Future Work

- Improve UI to incorporate multi-modal outputs, chat history etc.
- Train a specialized local LLM to achieve greater response accuracy and enhance context awareness. Use of approaches like transfer-learning.
- Scale and deploy on intra-networks to ensure high availability, ease of access.

# References

- [1] M. A. Kia, A. Garifullina, M. Kern, J. Chamberlain and S. Jameel, "Adaptable Closed-Domain Question Answering Using Contextualized CNN-Attention Models and Question Expansion," in IEEE Access, vol. 10, pp. 45080-45092, 2022, doi: 10.1109/ACCESS.2022.3170466.
- [2] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv preprint arXiv:2005.11401. Retrieved from <https://arxiv.org/abs/2005.11401>
- [3] Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. arXiv preprint arXiv:2104.08663. Retrieved from <https://arxiv.org/abs/2104.08663>
- [4] Tran, P., Nguyen, D., Tran, H.-A., Nguyen, T., & Tran, T. (2023). Building a Closed-Domain Question Answering System for a Low-Resource Language. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(3), 82. <https://doi.org/10.1145/3566123>



# Thank You!

Questions?