askMCQ

A RAG system for medical multiple choice question answering

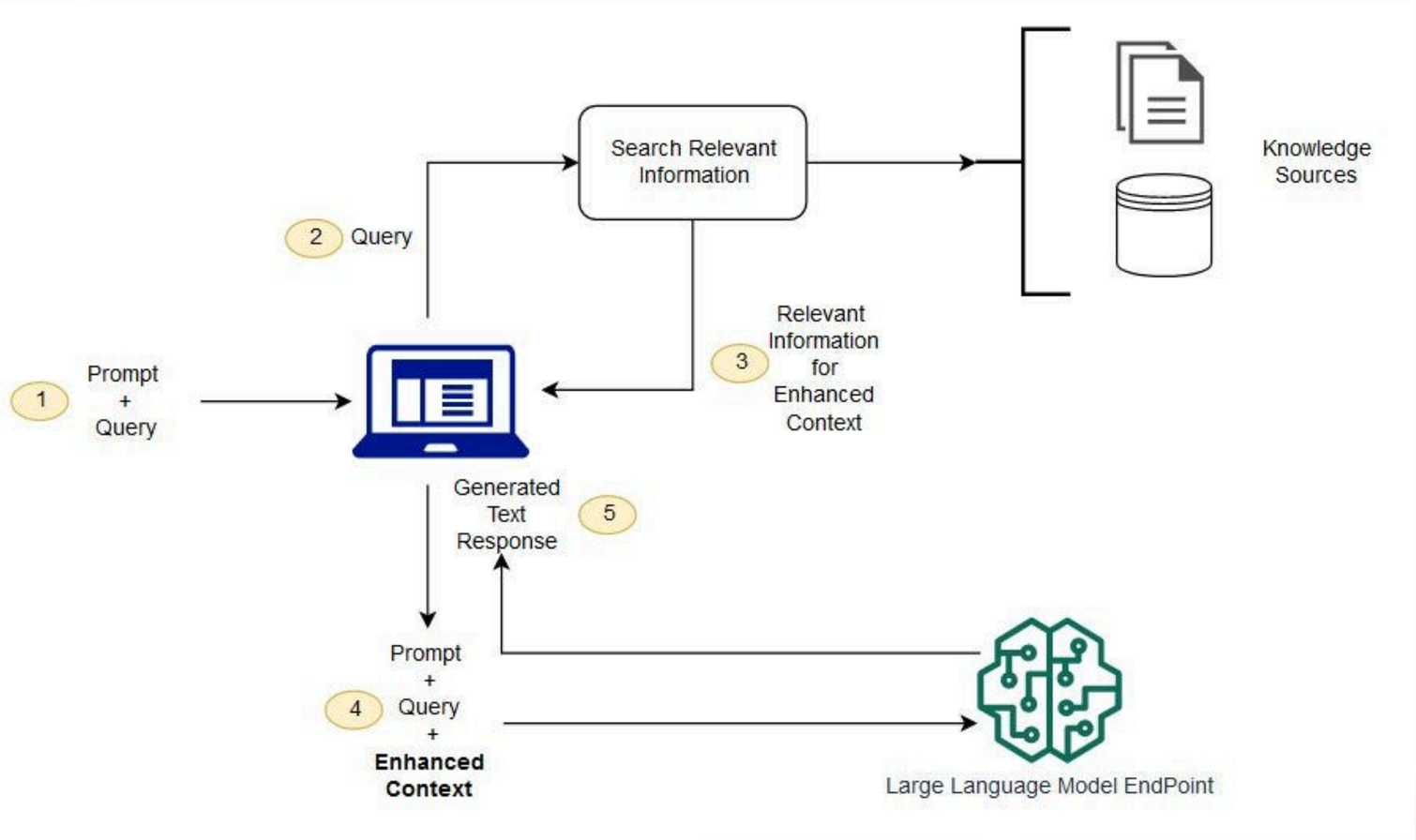
Anil Paudel

Objective

- Create an AI system for Medical case MCQ Answering sytem

- Focus on Reasoning and Retreival to get better answer

A Question Answering System



Challenges

- Establish proper knowledge base
- Optimize Retreival Performance
- Enchance LLM Reasoning for text generation

Creating Medical Context Database

Steps to create medical context database

Collection

Medical treatment guidelines are a set of recommendations for diagnosing, treating, and preventing diseases and conditions.

Diverse medical conditions

Acute Coronary Syndrome, Acne, Hematology, Hypothyrodism, Gonorrhoeae, Adult Sinusitis,Episodic Migraine Prevention, Galucoma, Alzehimer, Anxiety, OCD and PTSD, Chronic Kydney Diseasem Hepatitis C in Chronic kidney Disease, Rheumatoid Arthritis

Creating Medical Context Database

Challenges

Variable formats of guidelines

No Prior Domain Knowledge

Unstructured

Creating Medical Context Database

Steps to create medical context database

PreProcessing

Parsing: **llamaparse** (pdf to markdown)
better context understanding and better parsing

Chunking

Employed two variation of chunking

1. **MarkdownHeaderTextSplitter**
2. **RecursiveCharacterTextSplitter**

```
CHUNK_SIZE = 300  # Size of each split chunk
CHUNK_OVERLAP = 30  # Number of overlapping characters between chunks
```

Embedding

embedding model: OpenAI '**text-embedding-3-small**'

Vector Database:

chroma-db
persistence mode

Generating 100 MCQ test set

Steps to generate clinical scnario based question

Structure of question

Prompt to NotebookLM

You are an expert in medical guidelines and clinical reasoning. Use the context provided to answer the user's question accurately and justify your response with reasoning based on the guidelines.
User's Question: {question}
Retrieved Context: {context}
Available Options: {options}

Task: 1. Based on the context and the question, choose the most appropriate answer from the provided options.
2. Provide detailed reasoning for your choice using the retrieved context and guidelines.
Output Format:
{ {{"selected_answer"}}: "<Your chosen option>", "reasoning": "<Your detailed reasoning>", "is_correct": "<true/false>", "evaluation": "<Compare your reasoning with the ground truth reasoning and provide feedback>" }

Ground Truth for Evaluation: { "correct_answer_text": "{correct_answer_text}", "correct_answer_idx": {correct_answer_idx}, "ground_truth_reasoning": "{reasoning}" } the variables are {correct_answer_text}, {correct_answer_idx},{reasoning},{question} {context} and {prompt} other are not variable

Some hiccups with NotebookLM

Some questions had missing answers
The **ground truth reason, ground truth context** and the answer itself are not under strict evaluation

Token generation limit

Question format

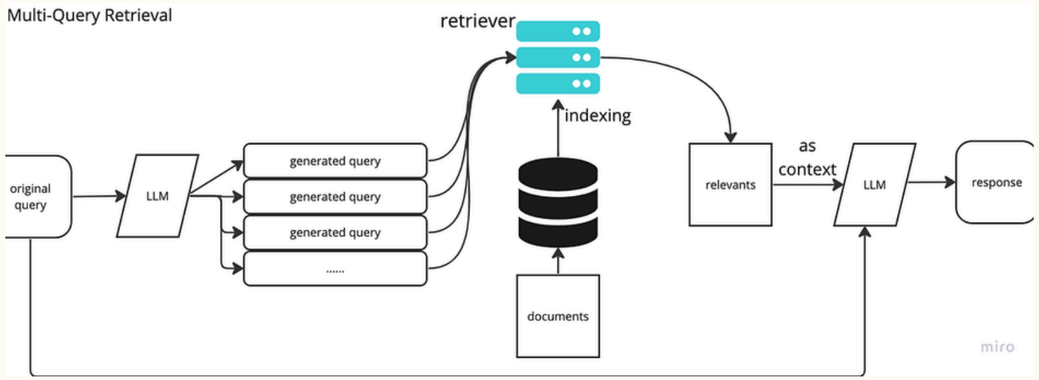
Some questions had missing answers
The **ground truth reason, ground truth context** and the answer itself are not under strict evaluation


```
{
    "question": "A 45-year-old female with chronic rhi
    "options": [
        "Prescribe a course of oral antibiotics.",
        "Recommend allergy testing and consider immuno
        "Perform nasal endoscopy to assess for polyps
        "Obtain a CT scan of the sinuses to assess for
    ],
    "correct_answer_text": "Perform nasal endoscopy to
    "correct_answer_idx": 2,
    "reasoning": "The patient has persistent CRS despi
    "ground_truth_context": "The clinician should conf
},
```

The Retrieval Component

LangChain Multi-Query Retrieval

- Enhances query generation by rephrasing or expanding user queries into multiple variations.
- Increases the likelihood of retrieving highly relevant documents, even for complex or ambiguous questions.



Similarity Search Retrieval:

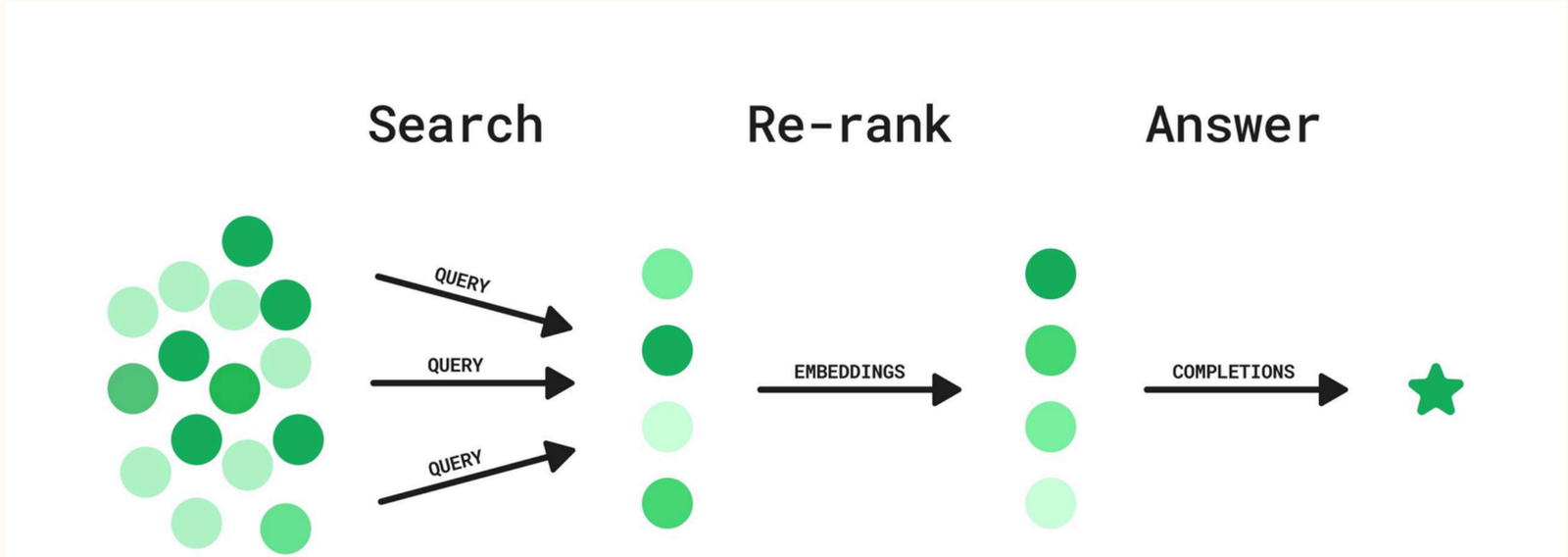
- Embeds queries and documents into a shared vector space.
- Finds and retrieves documents based on cosine similarity or other distance metrics, ensuring contextual relevance.

The Retrieval Component

Reranking Retrieval

- reranks the retrieved query using cross-encoder
- A Cross Encoder refines these results by deeply analyzing the semantic relationship between the query and the retrieved documents .

cross-encoder/ms-marco-MiniLM-L-6-v2



The Generation Component

A transformer-based model (gpt-4o-mini)

- a compact and efficient version of **GPT-4**
- capable of **understanding** complex **medical terms** and providing precise answers
- **cost effective** and **low computational** cost on my local machine

RAG Framework Integration:

Combine LangChain’s retrieval mechanism with generative models

Document Storage and Retrieval:

Similarity Search: Embeds queries and documents into a shared vector space for similarity-based matching.

Multi-Query Retrieval: Generates multiple query reformulations to retrieve diverse yet relevant documents.

Contextual Generation:
Document-to-Context Pipeline: Selects top-k retrieved documents and processes them to form input context for the generative model.

Generative Model (GPT-4o-Mini):
Generates responses using the provided contex

The Prompt Template

You are an expert at answering multiple-choice questions (MCQs). Your task is to analyze the provided context and select the correct answer (A-D) based exclusively on the information given. Follow these steps:

1. Context Analysis: Thoroughly read and extract key details from the context below.
2. Question & Options**: Identify the question's objective and evaluate each option against the context.
3. Answer Selection: Choose the option **directly supported by the context**, even if other options seem plausible externally.
4. Output Format: Return **only the letter (A-D)** as the answer, followed by a concise, context-grounded reasoning.

Context:
{context}

Question:
{question}

Options:
{options}

Response Format:
Correct Answer: [A/B/C/D]
Reasoning: [Step-by-step explanation using explicit context references. Avoid assumptions or external knowledge.])

The Prompt Template

You are an expert at solving multiple-choice questions through Chain-of-Thought reasoning. Analyze the question systematically and provide distinct reasoning paths.

Instructions
1. Context Analysis:
- Carefully read all provided context
- Identify key facts, relationships, and contradictions

2. Option Evaluation:
- For each option (A-D):
a) Check direct support in context
b) Identify conflicts with context
c) Note assumptions required

3. Reasoning Process:
- Consider at least 2 different interpretation angles
- Explicitly state if context is insufficient for any option
- Acknowledge and resolve ambiguities

4. Final Answer Selection:
- Choose*Format Requirements**
- Start with "Analysis:" for your reasoning
- Conclude with "Correct Answer: [X]" on final line
- Never use markdown or extra formatting

Post Processing of Response

RunnableMap

Instructing model to output JSON response, used **RunnableMap** to extract dynamically from response.

The response was not always JSON, **fine-tuning** needed so we moved on with **Simple Regex**

Used Simple Regex Search

The response was converted to plain text using **StrOutputParser** and matched with regex pattern for **Reason and Correct Answer choice** Extraction

Output

For Evaluation Purpose, the model response was combined with **retrieved context, ground truth reasoning**, **ground truth context**


```
{
  "question": "A 70-year-old patient who underwent percutaneous coronary intervention (PCI) with a
  "context": "|I|C-EO|In patients treated with DAPT after coronary stent implantation who must unde
  "ground_truth_context": "For patients with SIHD who are undergoing elective noncardiac surgery, i
  "model_answer": "C",
  "model_answer_idx": 2,
  "correct_answer_text": "Continue aspirin and stop clopidogrel 5 days prior to surgery.",
  "correct_answer_idx": 2.0,
  "model_reasoning": "The context states that for patients with stable ischemic heart disease (SIHD
  "reasoning_ground_truth": "For patients with stable ischemic heart disease who are undergoing ele
  "is_correct": true
},
```

Evaluation of System

Conclusion based evlauation

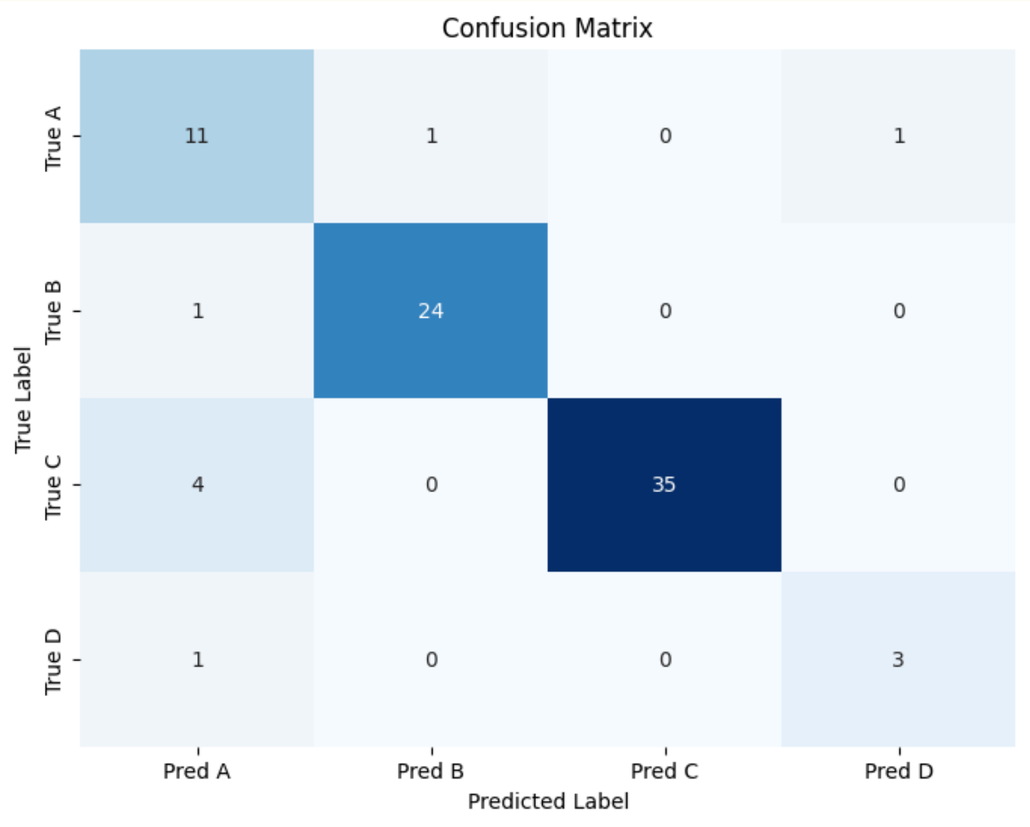
Evaluation Metrics used:

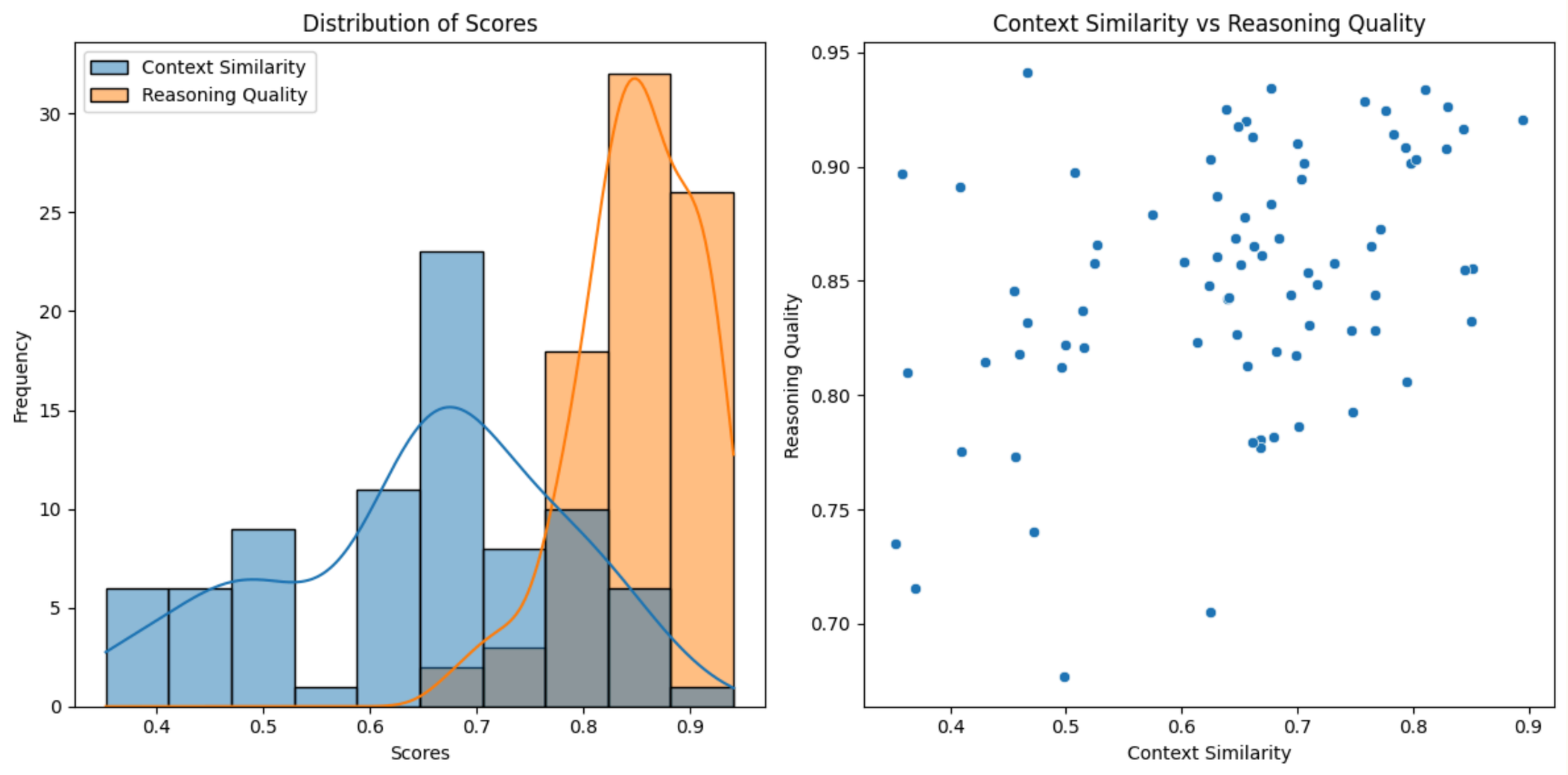
- Accuracy
- Confusion Matrix
- Similarity Score

Accuracy: 0.90

Total Samples: 81

Correct Predictions: 73





Cosine Similarity of the Embeddings with Ground Truth in both Context and Reasoning


Context Similarity vs Reasoning Quality
Reasoning Quality
Context Similarity
0.95
0.90
0.85
0.80
0.75
0.70
0.4
0.5
0.6
0.7
0.8
0.9

Experiments

Selection of Embeddings:

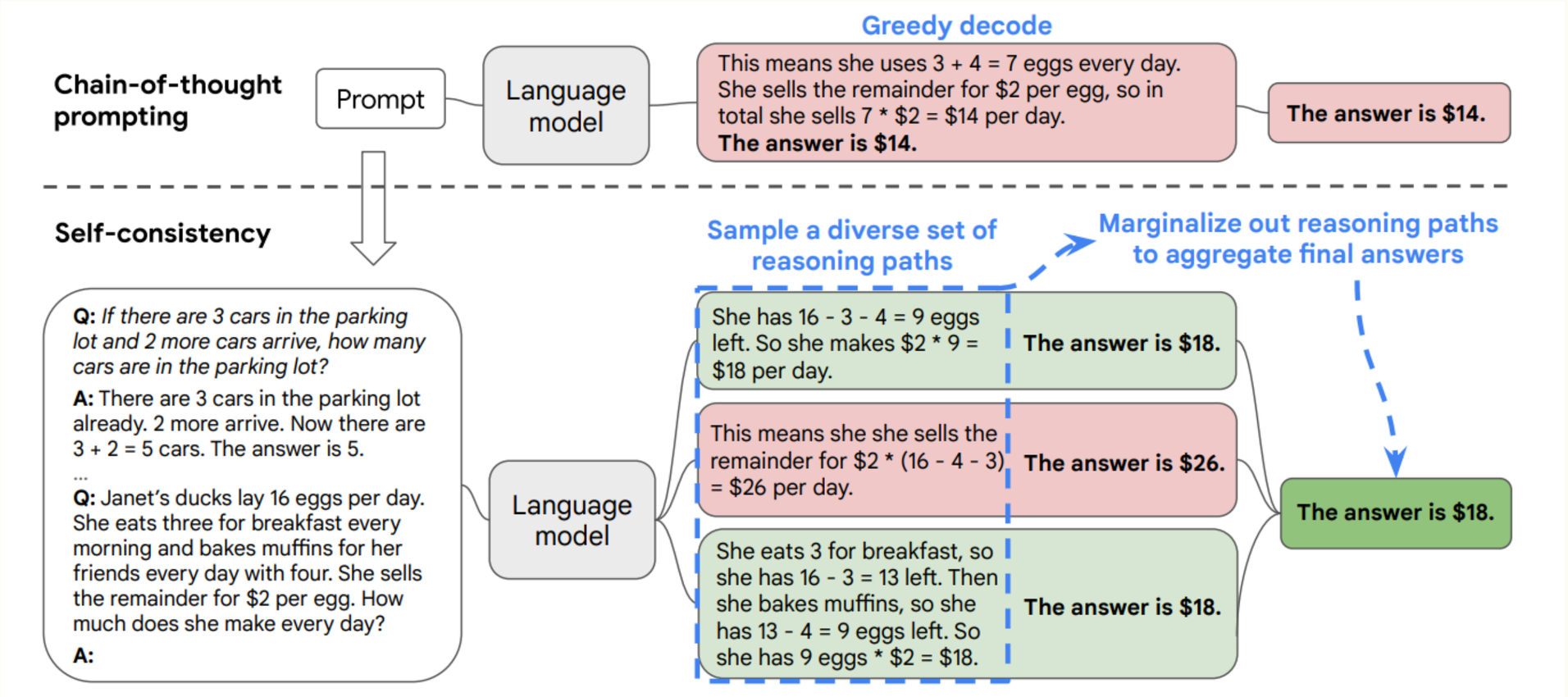
openai
text-embedding-3-small

Fine-Tuned Embedding Models for Medical Information Retreival
abhinand/MedEmbed-small-v0.1
abhinand/MedEmbed-base-v0.1

Selection of different retrieving techniques
Observed the response and context obtained from retrieving techniques lie: similarity search, multi query retrieval, reranker.

Generation Models: OpenAI gpt-4o-mini and gpt-4o

Zero shot CoT with Self Consistency: Tried zero shot CoT with majority voting for self consistency



Findings

Accuracy increased if queries as well as options are sent to retriever

Allowing model to take multiple reaoning path and selecting consistent answer result in improvements

Few shot prompting would allow the model to refine the apporach / decompose questions in even more fine way

gpt 4o model performs better but, sometimes doesn’t return expected format.

Learnings

Prompt makes difference Experimentation is necessary

Although the right answer was returned, prompt changes how the model responded

FewShotPrompting can result better performance and reasoning

Low domain expertise - Fully dependent on LLM reasoning

Data Preparation is the Key

Highly curated data with proper reasoning required to assess clinical case scenario

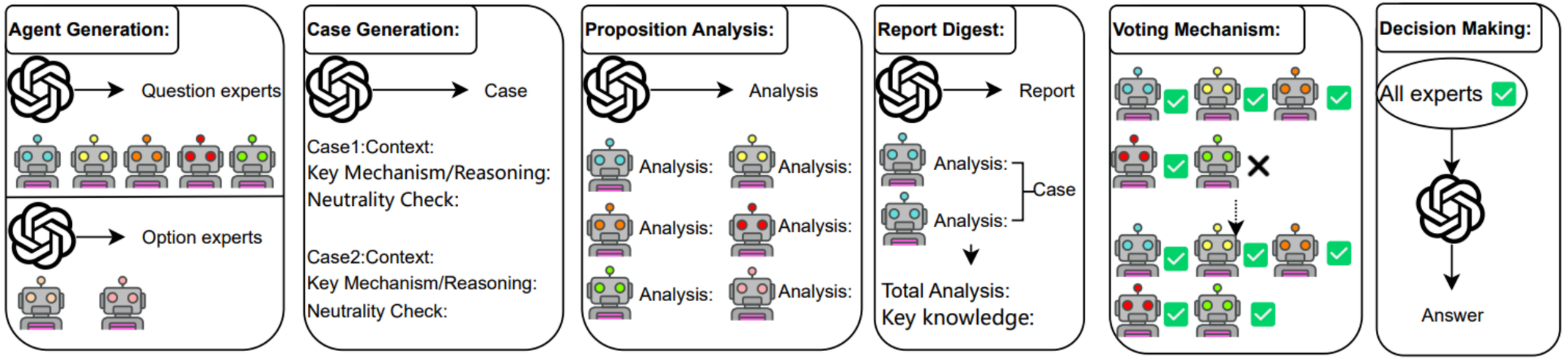
Better to use ;Medically Fine tuned LLMs

Med-PaLM 2 current best SOTA model on Medical QA

Knowledge Graph Integration would enhance retrieval with structured domain knowledge

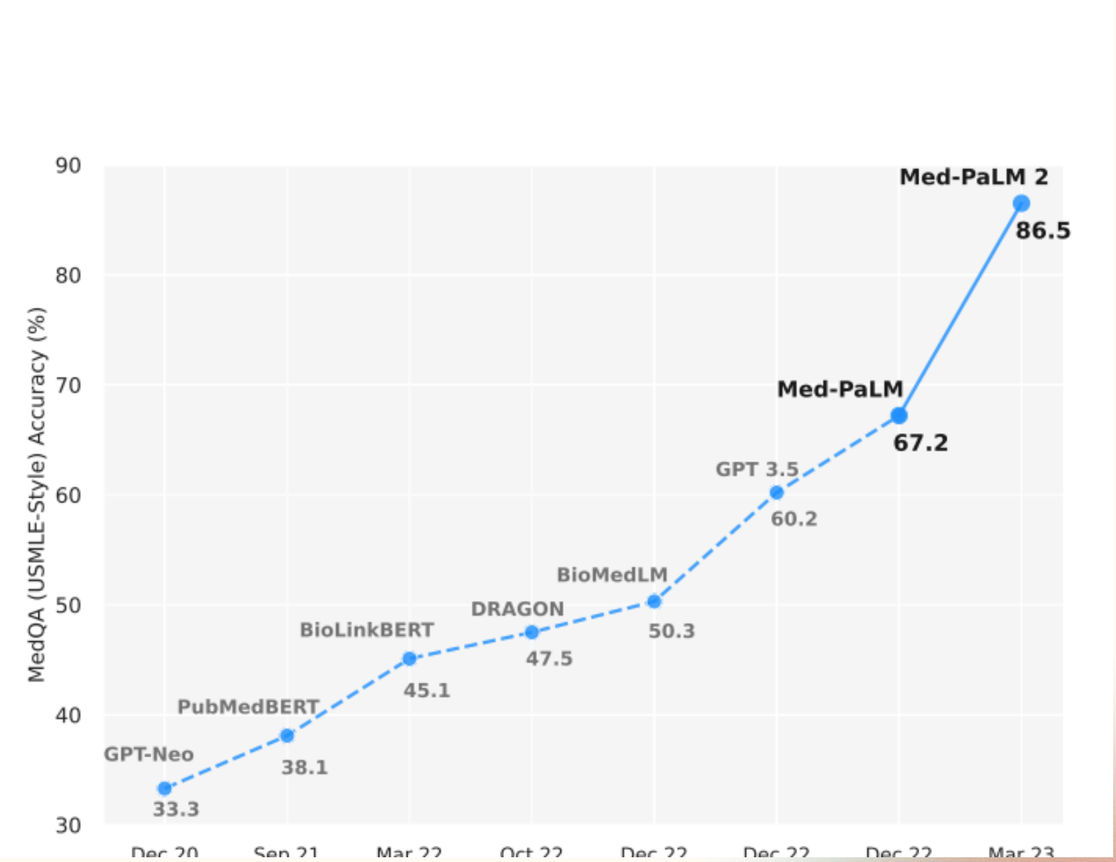
LLM-MedQA

Question:A 65-year-old male presents to the emergency department from his home complaining of dyspnea. He is alert and oriented. The following arterial blood gas readings are drawn: pH: 7.33 (Normal: 7.35-7.45), pCO2: 70 mmHg (Normal: 35-45 mmHg), HCO3 33 (Normal: 21-26 mEqL) Which of the following is most likely to have produced this patient condition?
Options:A) Panic attack. B) Mechanical ventilation. C) Diabetic ketoacidosis. D)Pulmonary embolus. E)Chronic obstructive bronchitis.



Ref:https://arxiv.org/abs/2501.05464

LLMs Performance in Medical QA



Knowledge Limitations:

Hallucination:

Context Sensitivity:

Ethical and Legal Concerns:

LLMs should not be solely relied upon for critical or high-stakes medical decisions

Assistive technology integrated into a robust pipeline with human oversight, domain-specific model and validation mechanisms.

Timeline of the Project

Learn Langchain Ecosystem	In order to utilize all the tools required, I spend first few days learning Langchain ecosystem
Build a complete QA pipeline	A very simple pipeline was built that can retrieve context and answer the MCQ
Context Database + Test MCQ generation	Used different resources to gather, preprocess and index context db, used NotebookLM for MCQ generation
Fully integrate into Pipeline	Built the system using the context database and LLM (GPT 4o mini) for Genration
Experimenting	Experimented Retrieval techniques, chunking techniques, prompt types
Evaluation and Visualsation	From the obtained response, evaluate the accuracy, context similarity and present in notebook/slides. Presentation Preparation

Thank You

Thank you for listening!