RoostAI: A University-Centered Chatbot

10.22.2024 Progress Update

Name: Vansh Nagpal

Major: CS/Math

Minor: DS

Role: ML Engineer

Name: Nitin Gupta

Major: CS

Minor: Math/DS

Role: ML Engineer

UNIVERSITY OF
South Carolina

Let’s Review Our Problem/Proposed Solution

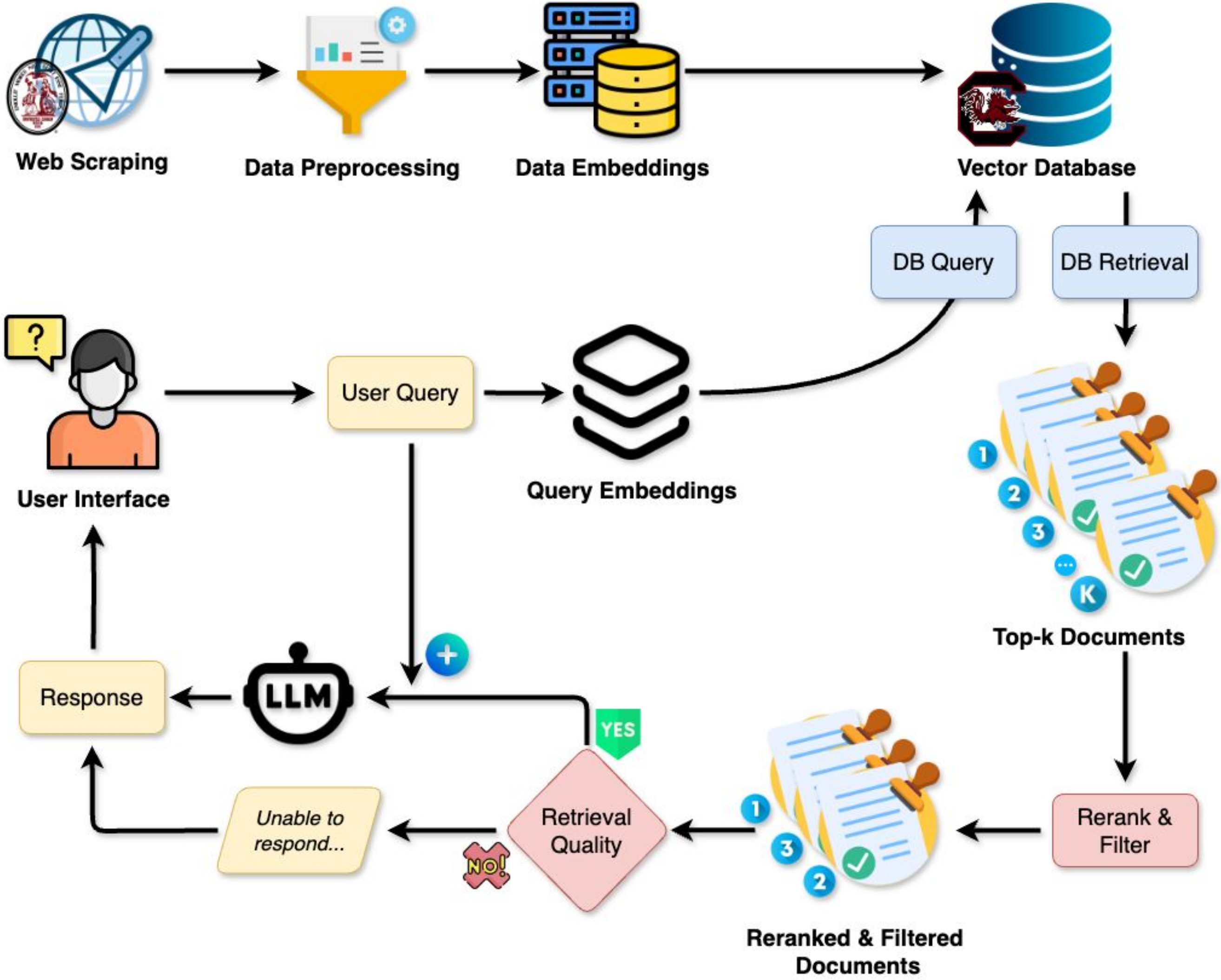
- USC lacks a comprehensive easy-to-use resource that informs them of campus related information.
- No AI chatbot exists to provide students with a reactive/interactive UI to quickly ask questions
- Project type: Implementation within a specific domain

Prototype Logo

The Solutio n

The RoostAI framework

updated

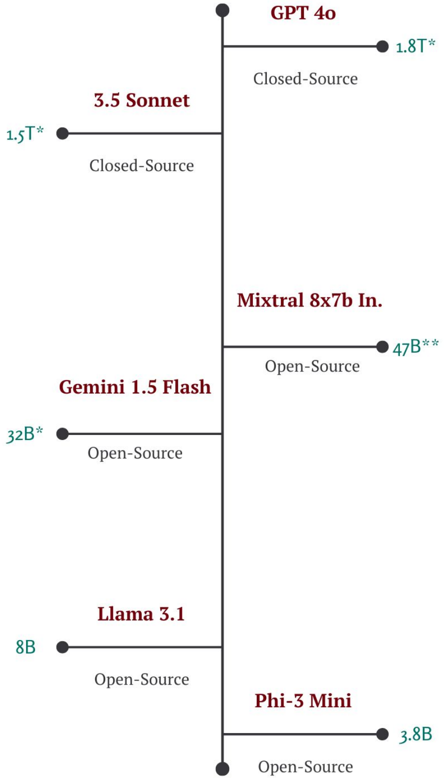


Evaluation

- **Numerical Analysis**
  - Latency and efficiency. *System*
  - Retrieval metrics (Mean Reciprocal Rank (MRR), Precision@k, Recall@k, etc.) *RAG*
  - Generation metrics (BLEU, ROUGE, Perplexity, etc.) *LLM*
- **Validation/Test Sets from USC FAQs**
  - Relevance, coherence, fluency, factual accuracy, etc.
- **User study to rate system**

LLMS CONSIDERED

Ordered by Parameter Count



GPT 4o

1.8T*

Closed-Source

3.5 Sonnet

1.5T*

Closed-Source

Mixtral 8x7b In.

47B**

Open-Source

Gemini 1.5 Flash

32B*

Open-Source

Llama 3.1

8B

Open-Source

Phi-3 Mini

3.8B

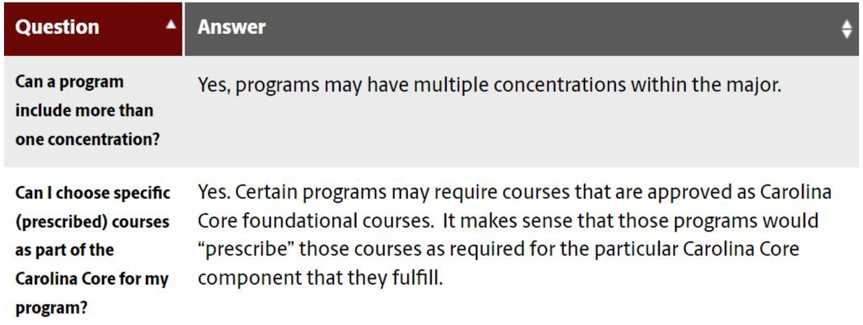
Open-Source

*Estimated parameter counts (unofficial).

** Total parameter count given;
however, not all parameters are used during inference.

LLM Experimentation on FAQs

- **Goal:** Narrow down candidates for our RAG response engine
- **Data: 40 USC FAQs** as our ground truth responses
- **Eval: Reference-based metrics** on LLM responses and ground truth

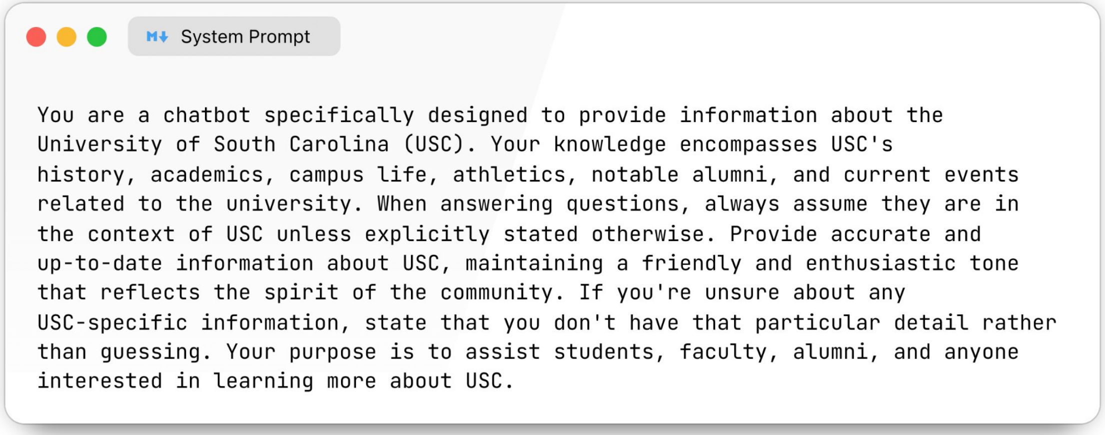


Question ▲	Answer ▲▼
Can a program include more than one concentration?	Yes, programs may have multiple concentrations within the major.
Can I choose specific (prescribed) courses as part of the Carolina Core for my program?	Yes. Certain programs may require courses that are approved as Carolina Core foundational courses. It makes sense that those programs would "prescribe" those courses as required for the particular Carolina Core component that they fulfill.

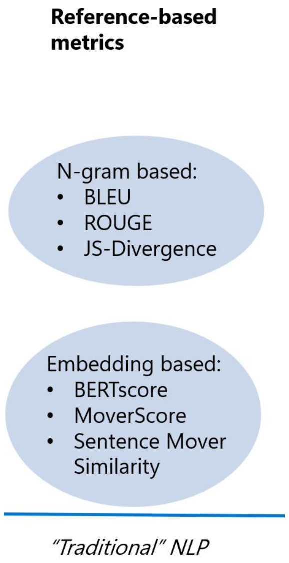
Metrics: *ROUGE*, *BLEU*, *Levenshtein distance*, *JS Divergence*, *Jaccard Similarity*, and *BERT Score*

Experimental Design

- **Independent variables:** FAQ
- **Control:** Metric being considered
- **Dependent variables:** LLM Metric score



You are a chatbot specifically designed to provide information about the University of South Carolina (USC). Your knowledge encompasses USC's history, academics, campus life, athletics, notable alumni, and current events related to the university. When answering questions, always assume they are in the context of USC unless explicitly stated otherwise. Provide accurate and up-to-date information about USC, maintaining a friendly and enthusiastic tone that reflects the spirit of the community. If you're unsure about any USC-specific information, state that you don't have that particular detail rather than guessing. Your purpose is to assist students, faculty, alumni, and anyone interested in learning more about USC.

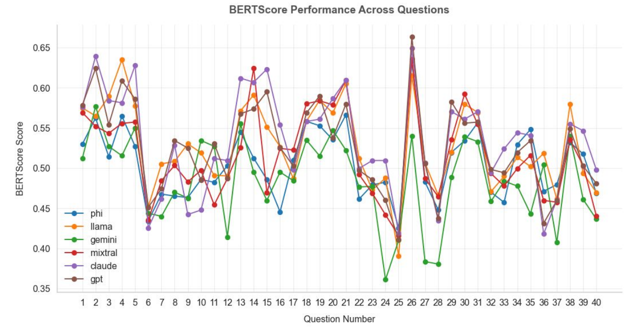

Reference-based metrics

N-gram based:
- BLEU
- ROUGE
- JS-Divergence

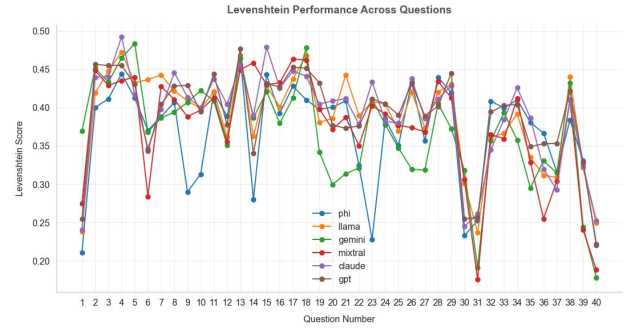
Embedding based:
- BERTscore
- MoverScore
- Sentence Mover Similarity

"Traditional" NLP

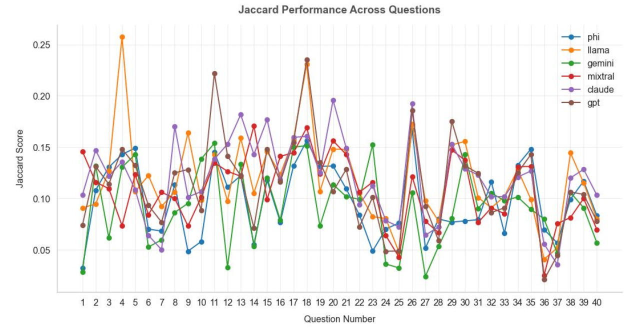
BERTScore Metric Comparison



Levenshtein Metric Comparison

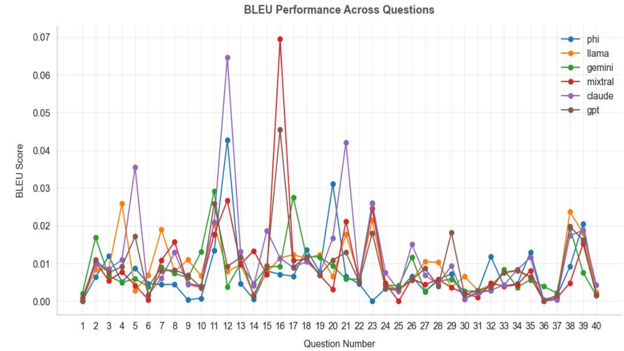


Jaccard Metric Comparison

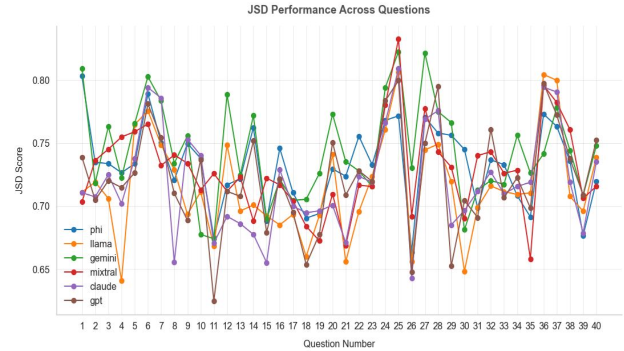


Evaluation Results

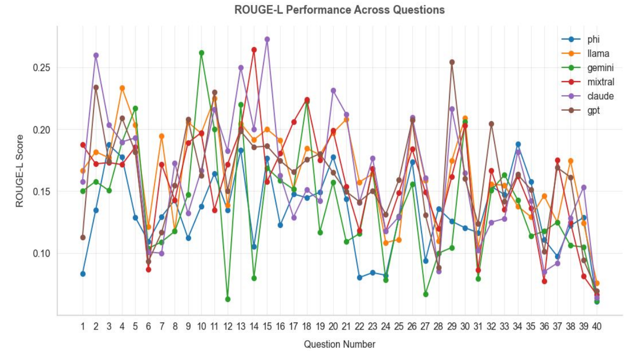
BLEU Metric Comparison



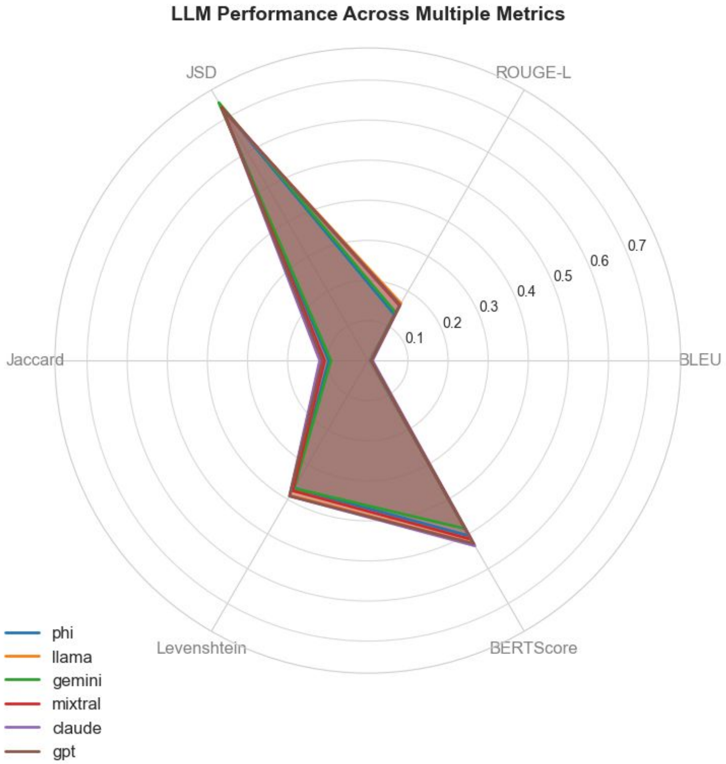
JSD Metric Comparison

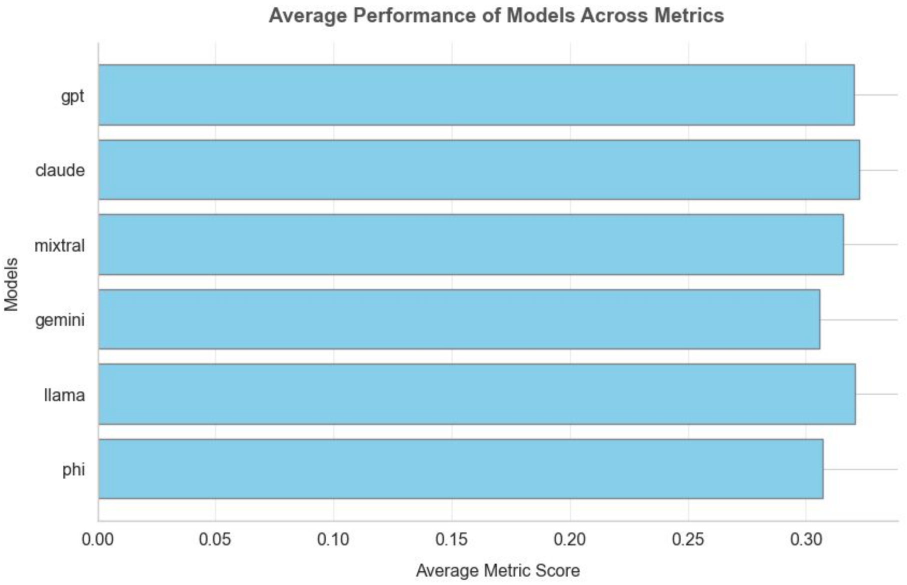


ROUGE-L Metric Comparison



Evaluation Results (cont.)





Discussion

- **Experimentation Insights:**
  - Metrics Might Be Inadequate for Fine-Grained Differentiation
  - Homogenization of Training Data
  - Current LLMs offer a solid baseline for handling FAQ-style queries
- **Next Experimental Steps:**
  - Evaluation of Non-FAQ Scenarios
  - Integrating Retrieval Mechanism

Other Updates

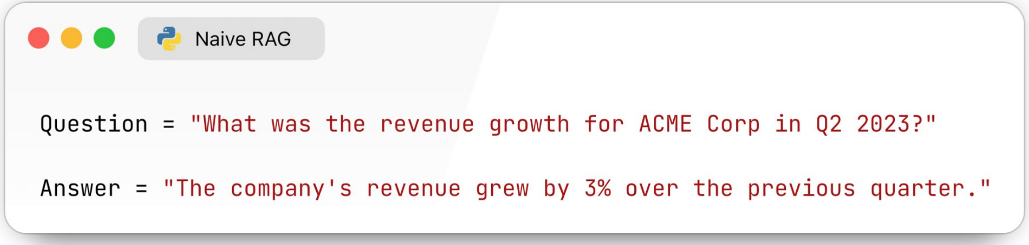
- Web scraping (*Beautiful Soup*, *Selenium*)
- Semantic Chunking of HTML Documents (*LlamaIndex*, *LangChain*)
- LLM-based Contextualization of Chunks (*Mixtral*, *Llama* or similar)
- Embedding of Contextualized Chunks w/ Encoders (*BERT* or similar)
- Creating of Vectorized DB (*ChromaDB*, *Pinecone*)
- User Application Creation (*Gradio*, *Streamlit*)

Chameleon

Completed; Yet To Be Completed

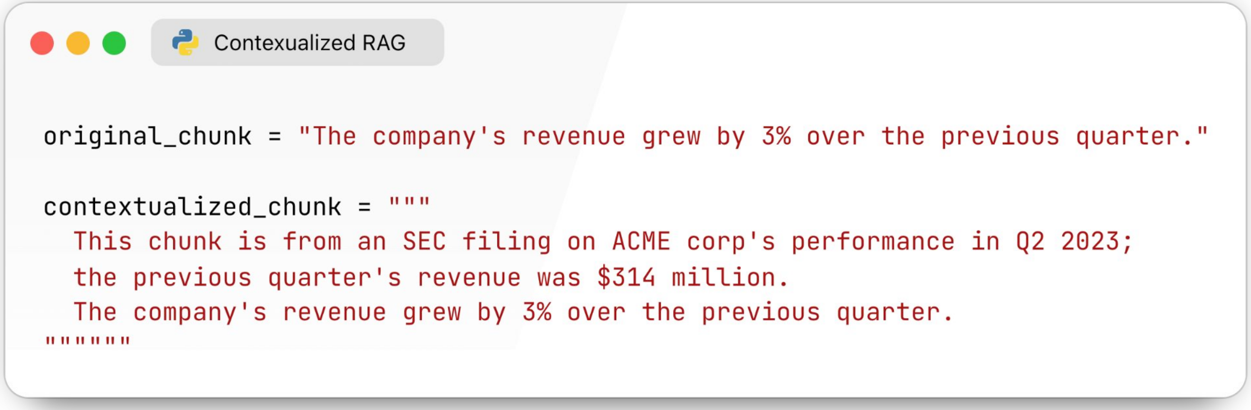
Backup Slides

Contextualized RAG



```
Question = "What was the revenue growth for ACME Corp in Q2 2023?"

Answer = "The company's revenue grew by 3% over the previous quarter."
```



```
original_chunk = "The company's revenue grew by 3% over the previous quarter."

contextualized_chunk = """
  This chunk is from an SEC filing on ACME corp's performance in Q2 2023;
  the previous quarter's revenue was $314 million.
  The company's revenue grew by 3% over the previous quarter.
""""""
```

Source:
https://www.anthropic.com/news/contextual-retrieval