Capstone Progress Report 1

Last updated: Feb 28, 2025

Stakeholder Names and Roles

In the table below, enter information for each team member, mentor, and sponsor. On the sponsor side, there may be several stakeholders including project manager, data contact, etc.

Stakeholder	**Role**
Bereket Tafesse	*Team Member*
Doruk Ozar	*Team Member*
Luke Napolitano	*Team Member*
Vishwanath Guruvayur	*Team Member*
Dr. Brian Wright	*Mentor*
Lucas McCabe	*Sponsor*
Dr. Brant Horio	*Sponsor*

Project Title: Multimodal RAG LLM in Education

Abstract:

The rapid advancement of Large Language Models (LLMs) is increasingly being considered for multimodal data processing. However, the effective retrieval of information from diverse data sources, for example, in educational contexts, still has many challenges. This capstone project explores the application of Retrieval-Augmented Generation (RAG) methods for multimodal data retrieval, focusing on a classroom setting where audio, video, and text data are made available. Our primary aim is to investigate methods that allow a foundational LLM(s), which have not been pre-trained on such multimodal data, to effectively search and retrieve relevant information across these different data types.

In this study, we will collaboratively design and execute experiments to determine the most effective approaches for multimodal retrieval, taking into account the unique modal characteristics of the data. For instance, we might explore the capability of these models to accurately infer the content of a lecture by analyzing and integrating inputs from audio

recordings, video footage, and textual transcripts. Additionally, we will examine the utility of various retrieval methods in delivering contextually appropriate and accurate information in response to user queries.

Throughout the discovery phase, we will collaboratively define a set of research questions that guide your experiments, such as assessing whether RAG methods can optimize retrieval accuracy or how these models perform under different multimodal input conditions. By the end of this project, we aim to demonstrate practical applications of multimodal retrieval in education, providing insights into the strengths and limitations of current approaches and paving the way for future advancements in this area, and other domains.

Outline of the Project

- *The Business purpose of the project*
    - *Improve independent individual learning outcomes (students using AI chatbot to learn when teacher isn't available)*
    - *Analyze use cases of Multimodal RAG*
- *Why the project is important?*
    - *Expand capabilities of LLMs using multimodal data*
    - *Developments in AI, is it really a helpful tool or is it a gimmick (learning outcomes)?*
- *Who are the stakeholders?*
    - *Students*
    - *Professors*
    - *AI researchers, LLM/NLP research*
- *Important Assumptions*
    - *Assumption that vectorization choice for all multimodal data is accurate*
    - *Retriever from document stores has no bias and all documents have correct access*
    - *Generation doesn't hallucinate too much and we can actually evaluate performance in the classroom as an educational tool*
- What is in scope / out of scope (as needed)
    - What is in scope?
    - How to evaluate these models, what do we consider successful implementation
    - What embedding models/vector representations are best for multimodal data?
    - Using backup plan, active RAG flows
    - Only using content from DS 3001, pdfs, lecture slides, hopefully moving into videos and audio to exactly pinpoint where relevant info is

- What is NOT in scope?
    - Website/package/chatbot front-end that we can put the RAG-bot into? Assuming that the UI is already built in the Text based RAG phase
    - Using the bot for multiple courses (Data Systems instead of Foundations of ML)

Success Criteria

Success criteria define the needs for a successful project. These should be identified with help from the sponsor. It is important to revisit these criteria with the sponsor as the project progresses to ensure alignment.

SC1	*Effective embedding of all multimodal data*
SC2	*FAISS scores achieve parity, embeddings are close*
SC3	*Retrieval is from the correct sources (e.g. pdfs find pages, video transcripts find min:sec interval)*
SC4	*Minimal hallucinations and corrective measures in instance of hallucination*
SC5	*Generalizable framework for other courses, can just apply RAG flow*

Data Assumptions and Limitations

Identifier	**Description**
A	*RAG is going to be better than newer chatbots with larger context windows and improved reasoning abilities*
A	*Vectorization will work the way it is supposed to with Multimodal Data*
A	*Retriever will grab from the right databases/vector stores*
A	*Minimal hallucination so we can evaluate classroom outcomes*
A	*The UI and Architecture for the Chatbot is already built.*
L	*Training RAG-bot for undergraduate machine learning class, nothing more or less advanced*

Summary of Data Processing, Data Aggregation

The dataset for our project is from the classroom setting of ML101 Class of Prof. Brian Wright
It includes – Lecture Slides, Lecture Audio Transcripts and Textbook PDFs

The Data is separated into two types based on modality – text and images.
The Images consist of lecture slide images of graphs and illustrations and textbook pdf images.

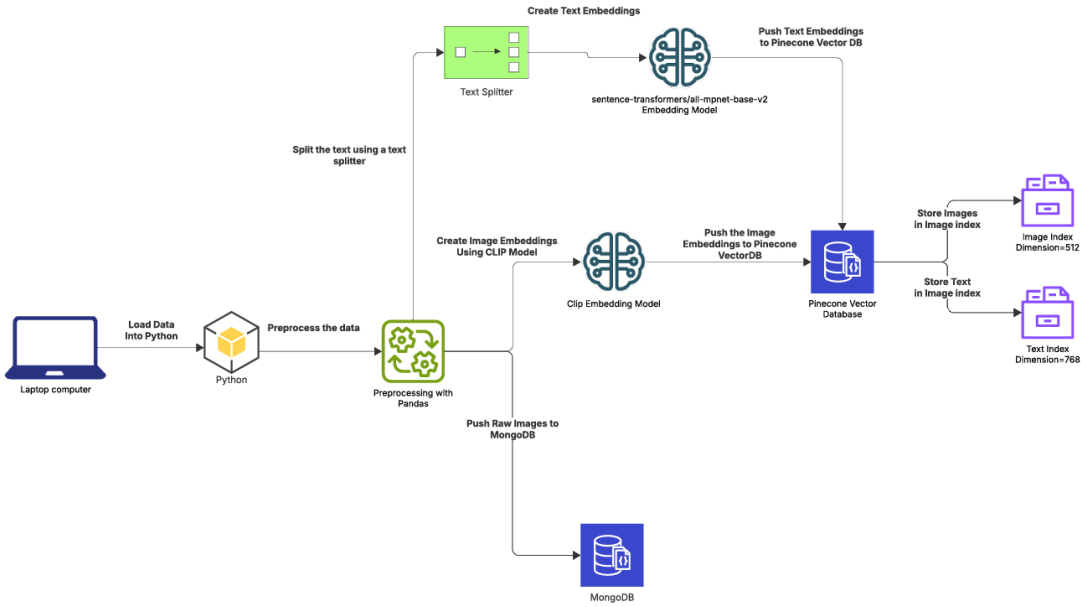
The Vectors are stored in Pinecone and the raw images are stored in MongoDB

The processing depends on the modality:

- Text – The textual content is vectorized using HuggingFace Sentence Transformers
    - Chunking – Size 1000, Overlap 100
    - Embedding using `sentence-transformers/all-mpnet-base-v2`
        - This embedding model is used because of its ability to work with language as well as scientific text
    - Stored in Pinecone with vector dimension 768

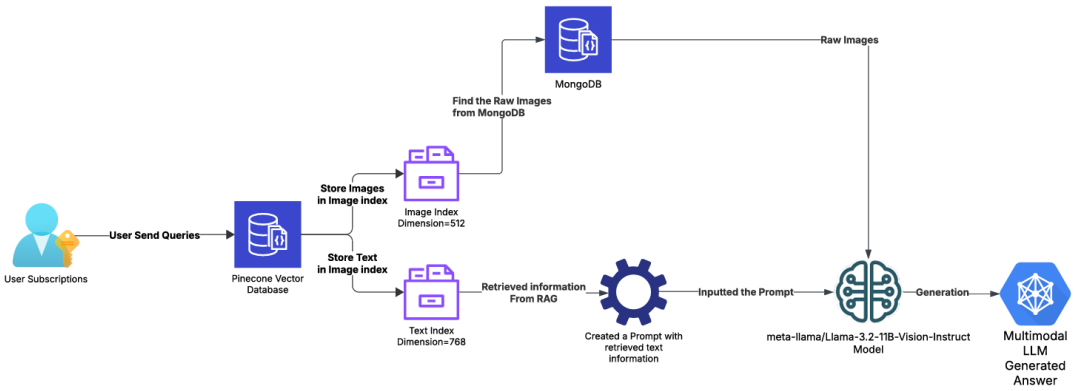
- Images – The images are embedded using the CLIP Transformer
    - Get the image embedding features using clip-vit-base-patch32
        - This embedding model is used because it is an OpenAI Pretrained transformer trained to maximize the similarity of image and text pairs via contrastive loss.
    - Stored in Pinecone with vector dimension 512
    - Raw Images stored in MongoDB for access

Storage Schema Diagram



Data Visualizations

This is a visualization of how the RAG Pipeline is currently designed.



Different Vector Stores for Images and Text and respective embeddings used to retrieve vectors for input query.

Summary of Modeling and Analysis

This project integrates a **Multimodal Retrieval-Augmented Generation (RAG) framework** that combines both textual and visual modalities for educational purposes. The main objective is to research about the inclusion of multimodality into the RAG framework and its benefit in Answer generation.

The aim is to evaluate and compare the performance of various Large Language Models (LLMs), including Llama 3.2, GPT-3.5, GPT-4, and potentially GPT-4o, across different retrieval-augmented generation (RAG) settings. The evaluation is structured into **three key stages**:

1. **Zero-Shot Performance:**
    - Assess baseline capabilities of each model without fine-tuning using **BLEU, ROUGE, and perplexity**.

2. **Tuned RAG Performance:**
    - Measure improvements in **faithfulness (RAGAS)**, **retrieval relevance (F1, MAP, MRR)**, and **answer relevance (ROUGE, METEOR)** after fine-tuning with a retrieval system.

3. **Multimodal RAG Performance:**
    - Extend evaluation to **multimodal inputs (text + images)**, incorporating additional factors affecting retrieval and answer relevance.

The main embedding and LLM models considered are:

EMBEDDINGS:

- **Text Embeddings**: HuggingFace Sentence Transformers (`all-mpnet-base-v2`)
    - This model is optimized for **semantic search and retrieval** tasks, making it suitable for handling scientific and educational text.
    - It produces embeddings of **dimension 768**, ensuring high contextual accuracy.
    - Works well with **lecture slides, textbook excerpts, and transcripts**.

- **Image Embeddings**: OpenAI CLIP (`clip-vit-base-patch32`)
    - CLIP is designed to **align images with text**, which is crucial for retrieving graphs, illustrations, and figures from lecture slides and textbooks.
    - Image embeddings are **dimension 512**, ensuring compatibility with **vector-based retrieval**.

LLMs chosen for comparison:

- Llama - meta-llama/Llama-3.2-11B-Vision-Instruct
- ChatGPT o1
- ChatGPT 4o
- ChatGPT o3-mini

Future Work Plan

Major Task Description	Date
Implement RAG Evaluation pipeline	Mid-March
Modularizing entire Codebase	Mid-March
Incorporate work from Prof. Brian's Lab	End-March
Write Research Paper	Mid-April

Potential Concerns [C] and Blockers [B]

Identifier	Description
Time [C]	*Do we have enough time to complete all aspects of our model and evaluation that we had planned*