Using Digital Textbook and Classroom Data to Explore Multimodal (Audio, Visual, & Textual) LLM Retrieval Techniques

Brian Wright[1], Vishwanath Guruvayur[1], Luke Napolitano[1], Doruk Ozar[1], Ali Rivera[1], Ananya Sai[1] and Bereket Tafesse[1]

[1]*University of Virginia, School of Data Science, Charlottesville, VA, United States*

Abstract

The use of digital content to support classroom learning is evolving rapidly. Retrieval Augmented Generation (RAG), as a approach to training LLMs, has emerged as a powerful framework to ground generation in trusted content. In the educational context this is materials sourced by professors/teachers for a specific courses. Although RAG systems traditionally rely on textual input, modern digital textbooks often includes a blend of modalities such as course slides, video lectures, and other interactive content containing both textual and visual information. In this project, we investigate the role of multimodal retrieval in an educational context using digital textbooks and other multimodal course data.

We embed and store textual and visual components from an undergraduate machine learning course into a vector database and use them to enhance chatbot responses. Through zero-shot RAG experiments and evaluation using RAGAS metrics such as Context Recall, Faithfulness and Factual Correctness, we examine how supplementing text with images impacts retrieval and response quality. Our findings show that multimodal input significantly improves factual correctness for complex or specific queries, although excessive image inclusion may reduce performance. Conversely, image inclusion does not provide gains on more generic questions. We propose an agent-based RAG system that dynamically selects relevant vectors based on query specificity.

Keywords

LLMs, Chatbot, RAG, Machine Learning

1. Introduction

Recent advances in LLM have catalyzed interest in the application of generative models to educational tools. However, standard LLMs lack awareness of the multimodal data prevalent in classrooms. This includes interactive elements of textbooks, slide visuals and lecture recordings. One promising approach to addressing this limitation is Retrieval Augmented Generation (RAG), which enables models to incorporate external knowledge into the generative process.

Rather than relying solely on pre-trained outputs, RAG retrieves relevant documents from an external corpus based on a user's query. In an educational context this includes material sourced or generated by professors/teachers. These augmented documents can be used to aid in the response generation process. This approach could prove to be particularly valuable in educational settings, where a chatbot can generate responses that align closely with the specific content and instructional level of any given course. Although RAG typically relies on text-based retrieval, we explore its extension to include both text and image embeddings from digital educational materials.

This is an exploratory study designed to position deeper understand on potential technical approaches for developing an LLM based Intelligent Assistant to support students with a focus on self-regulated learning. Our goal was to first understand the utility of using a RAG based approach with data from open source textbooks and lecture materials. This was followed by a exploration on whether the incorporation

The 26th International Conference on Artificial Intelligence in Education

bw2zd@virginia.edu (B. Wright); vish@virginia.edu (V. Guruvayur); ljn5yms@virginia.edu (L. Napolitano); bcp8dm@virginia.edu (D. Ozar); wat6sv@virginia.edu (A. Rivera); wxqr8dj@virginia.edu (A. Sai); uqs3dq@virginia.edu (B. Tafesse)
https://datascience.virginia.edu/people/brian-wright (B. Wright)
0000-0002-7836-8273 (B. Wright)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

of visual content enhances the generation of educational response and under what circumstances it may hinder it. The RAG-bot is specifically designed for a undergraduate course, Foundations of Machine Learning, taught at the University of Virginia School of Data Science.

2. Background

During the last decade the inclusion of AI driven education tools has increased dramatically resulting in the maturity of a new field; Artificial Intelligence in Education (AIEd) [9]. The rise in the presence of AI in global society and emersion into our everyday lives has not only produced the development of additional educational tools but has also driven the need for the creation of a new literacy. Growing out of Data Literacy [3], AI Literacy is maturing to the point of being referenced in educational programs and research Long and Magerko [7]. In addition, the belief that AI has the potential to continue to transform how we communicate, consume information, learn, and interact in society seems like a forgone conclusion. Consequently, the need to measure the effectiveness of teaching methods in pursuit of AI literacy is currently high with additional research still being needed. Ouyang and co-authors in constructing a AI literacy framework through a meta-analysis of papers spanning several disciplines made note of this point. The authors further suggest this is especially true of courses in Data Science oriented programs designed to teach AI fundamentals, that often pull students from a variety of backgrounds [11].

The nature of how AI driven tools get incorporated into higher education occurs at essentially three levels; instruction/service, learning, and administration, [4]. Instruction/service-oriented can be seen as tools that help instructors grade assignments, facilitate students in choosing courses or identifying university resources but do not directly aid in knowledge growth. Learning oriented is focused mostly on classroom applications with the goal of helping students achieve learning outcomes. This may include tutoring, providing learning materials, facilitating students self-guided learning or intelligent assistants that have been tailored to course content [2, 5]. This category could also include general Large Language Models that aid in answering student questions or in the case of Data Science or Computer Science generated code. Administrative tools are geared toward educational staff or professionals that function out of direct line of sight of students. These could be anything from business intelligence systems for financial analyses or application tools that help aid in the admissions processes.

This project focuses on the learning level by exploring the creation of a multimodal chatbot to help students in a specific course. The multimodal nature of the approach is a growing research area, but one that requires more attention [8]. The follow on work will not only present the tool but give students a understanding of how it is trained and opportunities to augment with new data throughout the course. Thus touching the previous referenced ideas of facilitating AI literacy. This also allows for a active learning approach known to be productive for learning in STEM environments [1].

The original RAG framework [6] introduced a method of augmenting LLM output with external documents. Follow-up research has explored knowledge-grounded dialogue, domain-specific retrieval, and image-text fusion models like CLIP [10]. Our work draws from these threads, but focuses on integrating image and text embeddings within RAG for a specific instructional context, aligning with efforts in educational NLP and multimodal LLMs.

3. Methodology

3.1. Data Sources

We curated multimodal data from DS3001: Foundations of Machine Learning, including:

- Lecture slides (text + images)
- Lecture audio transcripts (text)
- Open Source ML Textbooks (text + images)
- Open Source ML papers (text + images)

Images and textual content were extracted and segregated from lecture slides, machine learning research papers and textbooks in PDF format. Additionally, lecture videos' audio recordings were transcribed into text using a speech-to-text conversion tool and incorporated as part of the textual dataset.

3.2. Embedding Details

Textual Data: Chunks (1500 tokens, 100-token overlap) were embedded using SentenceTransformer `all-mpnet-base-v2` and stored in Pinecone DB (dim=768).
Visual Data: Images were embedded using OpenAI CLIP (`clip-vit-base-patch32`, dim=512) and indexed in Pinecone Vector Database with filenames stored as metadata. Then the binary representations of the raw images are stored in MongoDB. Upon completing the cosine similarity search, the filenames of the most relevant images are retrieved and used to fetch the corresponding raw images from the database.

3.3. Experiment Design

We tested multiple RAG configurations:

- **Zero-shot LLMs**: Not including the RAG component
- **Text Only RAG (10 text vectors)**: Using only text retrieved from our vector DB
- **Text + Image (5 text + 5 image vectors)**: Text with less relevant vectors replaced by top images
- **Text + Image (10 text + 10 image vectors)**: Addition of more visual information along with base textual information

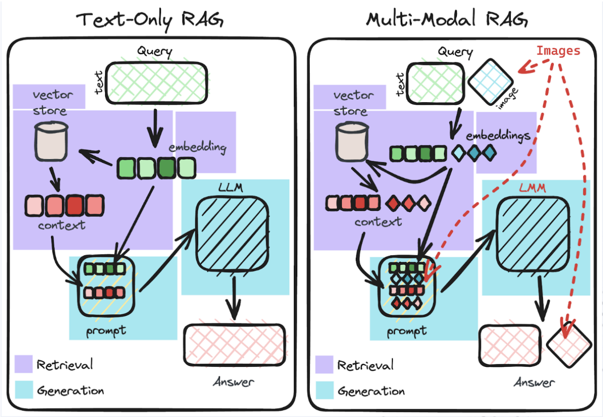

Figure 1: Visualization of how multimodality differs from text-only RAG

To properly visualize the performance of these experiments, we needed to compare their results to a zero-shot LLM using the most recent version of ChatGPT against the same questions. For fidelity, we created a multimodal information based dataset of 30 data science questions and answers to evaluate the performance of the models.

These questions ranged from generic data science questions like "What is K-Nearest Neighbors?" to highly domain specific questions for testing multimodal performance like "Explain Brian's Machine Learning Lifecycle." With a wide range of questions at our disposal, we could see how differing model types generalize to broad concepts and also validate how they perform on specific images from class content.

3.4. Evaluation

We used the evaluation package RAGAS to obtain metrics for our models:

- **Context Recall**: fraction of relevant documents retrieved (Context)
- **Faithfulness**: factual support in retrieved context (Query)
- **Factual Correctness**: overall factual alignment with the answer (Response)

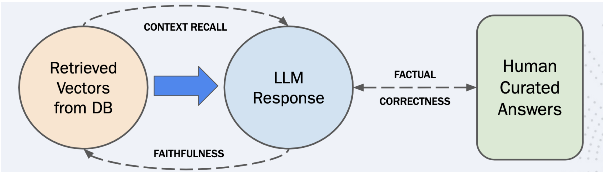

Figure 2: Visualization of how evaluation metrics interact with each component of the RAG model

These metrics were chosen to target three critical components of Retrieval Augmented Generation for evaluation. With RAG, we want to evaluate how the model performs in terms of the Context, the Query, and the Response, as seen in the list above each aligns with one of these components. From the query, we need to ensure that retrieved context is relevant to the query at hand. From the context, we need to ensure that the generated response aligns with the retrieved information. From the response, we need to ensure that the response actually answers the query we began with. Finding evaluation metrics for these aspect proved crucial to executing the experimental design of the study. Testing was bootstrapped using 50 diverse questions run 10 times allowing for the capture of variance in the evaluation metrics.

4. Experiment Results

We evaluated the effect of incorporating images into the Retrieval-Augmented Generation (RAG) workflow for educational question answering using a multimodal LLM (GPT-4.1 Nano). Our goal was to assess whether visual content improves response quality and contextual grounding, especially across question types of varying specificity.

4.1. Experimental Setup

We tested two configurations for image inclusion:

- **Text + Image (10T + 10I)**: Adds 10 image vectors to the 10 retrieved text vectors, preserving all textual context while layering on visual information.
- **Balanced Swap (5T + 5I)**: Replaces the bottom 5 text vectors with the top 5 image vectors, maintaining the same number of total context inputs but altering the text-image ratio.

Both configurations were evaluated on a curated dataset of generic and specific questions derived from course materials. We compared these against a Text-Only RAG baseline and a Zero-Shot (no retrieval) setting. Evaluation was based on three key metrics: *Context Recall*, *Faithfulness*, and *Factual Correctness*.

4.2. Generic Questions

For generic questions, in terms of the first goal of better understanding the variances between generic LLM versus a RAG system, it appears no real differences are present as the confidence intervals have significant overlaps for text only and text + images (10+10) approaches when compared to the zero shot model. However, the model does get significanyly worse in the text + images (5+5) approach.

For the RAG specific measures we observed that the 10T + 10I configuration modestly improved Context Recall compared to the Text-Only baseline. The 5T + 5I setup led to a more significant increase in recall, suggesting that adding well-ranked images can improve retrieval relevance.

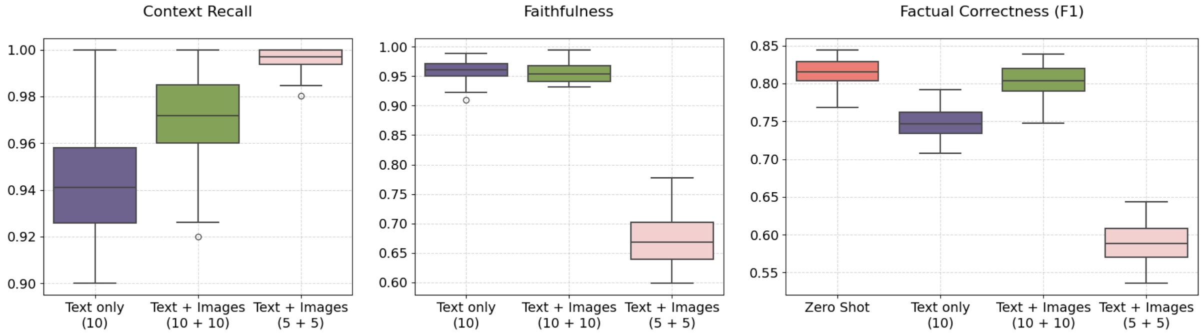

Figure 3: Performance Metrics on Generic Questions

However, this improvement came at a cost. Faithfulness and Factual Correctness declined in the 5T + 5I setup, likely due to the removal of text content that the LLM relied on for broader context and coherence. This tradeoff implies that generic questions—often answerable via general textual knowledge—benefit more from rich text contexts than from visual augmentation.

Summary: Image inclusion boosts Context Recall, but replacing even marginally relevant text hurts Faithfulness and Factual Correctness in generic settings. Retaining broader textual context is crucial for accurate and coherent answers, thus it might not be worth including images in all scenarios. Although we observed no differences between zero shot and RAG model metrics it is worth noting that using a RAG approach allows for the content to be easily updated and since performance was not worsened we would recommend this approach.

4.3. Specific Questions

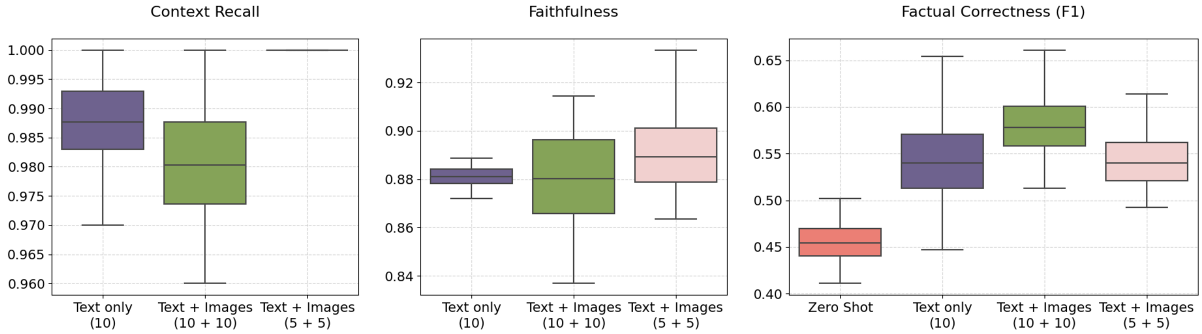

Figure 4: Performance Metrics on Specific Questions

For specific queries it did appear that the addition of the RAG based system added to the content being provided to users. This was again determined by analyzing the before and after effects of including and not including the RAG approach (zero shot) using factual correctness. The bootstrapped confidence intervals are either completely separate (Text + Images 10 + 10) or just slightly overlapping (Text + Images 5 + 5) for text + image approaches. Interestingly, this was not the case for the text only RAG approach. This suggest that the inclusion of images on more specific questions has a significant positive effect on the factual correctness of the LLM. Moreover, the RAG system allows for the tracking of where content is getting pulled from inside the vector database to supplement the generation of responses. Which could allow for a deeper level of understanding of relevant content as it relates to student questions.

For the RAG specific measures adding 10 images on top of 10 text vectors slightly reduced Context Recall, likely due to visual noise introduced by less relevant images. However, the 5T + 5I configuration achieved perfect recall consistently, showing that highly ranked visual content can provide strong contextual grounding for specialized queries adding further support for text and images on specific questions.

As compared to the generic questions faithfulness and factual correctness remained stable or slightly improved in both multimodal settings for specific questions. This suggests that relevant visual content supports accurate generation without undermining the consistency of the LLM responses.

Summary: For specific questions, selectively replacing lower-ranked text vectors with relevant images improves retrieval and enhances response quality. Excessive image inclusion, however, may distract the model. Overall, including images significantly improves results for specific questions when compared to zero shot models.

5. Conclusion and Future Work

This study explored the impact of multimodal retrieval, specifically the integration of image vectors within a Retrieval-Augmented Generation (RAG) framework for educational applications. Our experiments demonstrated that visual content, when selectively incorporated, can enhance the informativeness, contextual relevance, and response accuracy of a multimodal LLM, especially for conceptually dense.

However, our results also show that a fixed or naive strategy for image inclusion is suboptimal. In the 10T + 10I setup, the inclusion of excessive visual information led to performance degradation in certain metrics, particularly Faithfulness and Factual Correctness. These findings underscore the importance of context curation and relevance filtering in multimodal systems.

Future work will focus on developing dynamic, adaptive strategies to optimize retrieval and improve LLM responses. Key directions include:

- Designing an **agentic RAG selector** that adjusts the mix of text and image vectors based on real-time query specificity analysis.
- Exploring **semantic clustering and alignment** across modalities to better group and rank context vectors.
- Enhancing **evaluation efficiency** through smarter sampling, reproducible scoring pipelines, and reduced compute requirements.
- **Knowledge Graph based RAG** would work very well on this corpus of data as observed from the PCA Analysis of Clustered Text Vectors.

These improvements aim to support the development of intelligent, multimodal RAG systems that dynamically tailor context inputs—maximizing educational value and improving user engagement in classroom and self-guided learning environments.

References

[1] José Rafael Aguilar-Mejía et al. "Design and Use of a Chatbot for Learning Selected Topics of Physics". en. In: *Technology-Enabled Innovations in Education*. Ed. by Samira Hosseini et al. Singapore: Springer Nature, 2022, pp. 175–188. ISBN: 978-981-19-3383-7. DOI: 10.1007/978-981-19-3383-7_13.

[2] Vincent Aleven et al. "Help Helps, But Only So Much: Research on Help Seeking with Intelligent Tutoring Systems". en. In: *International Journal of Artificial Intelligence in Education* 26.1 (Mar. 2016), pp. 205–223. ISSN: 1560-4292, 1560-4306. DOI: 10.1007/s40593-015-0089-1. URL: http://link.springer.com/10.1007/s40593-015-0089-1 (visited on 05/21/2025).

[3] F. Javier Calzada-Prado and Miguel Marzal. "Incorporating Data Literacy into Information Literacy Programs: Core Competencies and Contents". In: *Libri* 63 (June 2013). DOI: 10.1515/libri-2013-0010.

[4] Maud Chassignol et al. "Artificial Intelligence trends in education: a narrative overview". In: *Procedia Computer Science*. 7th International Young Scientists Conference on Computational Science, YSC2018, 02-06 July2018, Heraklion, Greece 136 (Jan. 2018), pp. 16–24. ISSN: 1877-0509. DOI: 10.1016/j.procs.2018.08.233. URL: https://www.sciencedirect.com/science/article/pii/S1877050918315382 (visited on 03/28/2024).

[5] Lijia Chen, Pingping Chen, and Zhijian Lin. "Artificial Intelligence in Education: A Review". In: *IEEE Access* 8 (2020). Conference Name: IEEE Access, pp. 75264–75278. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.2988510. URL: https://ieeexplore.ieee.org/document/9069875 (visited on 03/29/2024).

[6] Patrick Lewis et al. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. arXiv:2005.11401 [cs]. Apr. 2021. DOI: 10.48550/arXiv.2005.11401. URL: http://arxiv.org/abs/2005.11401 (visited on 05/03/2024).

[7] Duri Long and Brian Magerko. "What is AI Literacy? Competencies and Design Considerations". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. New York, NY, USA: Association for Computing Machinery, Apr. 2020, pp. 1–16. ISBN: 978-1-4503-6708-0. DOI: 10.1145/3313831.3376727. URL: https://doi.org/10.1145/3313831.3376727 (visited on 03/27/2024).

[8] Mehrnoush Mohammadi et al. "Artificial Intelligence in Multimodal Learning Analytics: A Systematic Literature Review". In: *Computers and Education: Artificial Intelligence* (May 2025), p. 100426. ISSN: 2666-920X. DOI: 10.1016/j.caeai.2025.100426. URL: https://www.sciencedirect.com/science/article/pii/S2666920X25000669 (visited on 05/22/2025).

[9] Fan Ouyang and Pengcheng Jiao. "Artificial Intelligence in Education: The Three Paradigms". In: *Computers and Education: Artificial Intelligence* 2 (Apr. 2021), p. 100020. DOI: 10.1016/j.caeai.2021.100020.

[10] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020 [cs.CV]. URL: https://arxiv.org/abs/2103.00020.

[11] Elisabeth Sulmont, Elizabeth Patitsas, and Jeremy R. Cooperstock. "Can You Teach Me To Machine Learn?" In: *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*. SIGCSE '19. New York, NY, USA: Association for Computing Machinery, Feb. 2019, pp. 948–954. ISBN: 978-1-4503-5890-3. DOI: 10.1145/3287324.3287392. URL: https://dl.acm.org/doi/10.1145/3287324.3287392 (visited on 03/28/2024).

A. Appendix

A.1. Vector Store Visualization

This is a live link to an example of how questions and documents are embedded in our vector store. The most semantically similar documents used in the response are highlighted in purple and green. https://msds-capstone-project.github.io/MultiModalRAGViz/

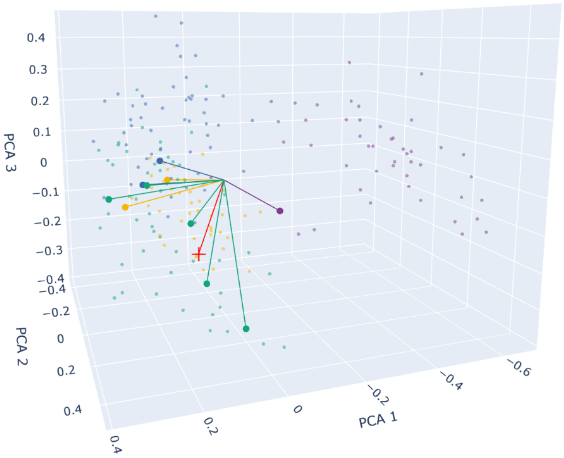

Figure 5: 3D Plot of PCA from 768 Dimension Text Vectors

A.2. Evaluation Metrics

These are the evaluation metrics calculated via 10 Bootstrapped sampling rounds of 50 queries each.

Table 1
Generic Questions Evaluation Metrics

Context Recall

Model	Mean	Std Dev
Text-Only	0.945	0.032
Text+Images 10+10	0.975	0.024
Text+Images 5+5	0.997	0.006

Faithfulness

Model	Mean	Std Dev
Text-Only	0.963	0.020
Text+Images 10+10	0.956	0.026
Text+Images 5+5	0.676	0.062

Factual Correctness (F1)

Model	Mean	Std Dev
ZeroShot	0.818	0.025
Text-Only	0.750	0.027
Text+Images 10+10	0.807	0.029
Text+Images 5+5	0.593	0.037

Table 2
Specific Questions Evaluation Metrics

Context Recall

Model	Mean	Std Dev
Text-Only	0.988	0.009
Text+Images 10+10	0.982	0.013
Text+Images 5+5	1.000	0.000

Faithfulness

Model	Mean	Std Dev
Text-Only	0.881	0.005
Text+Images 10+10	0.883	0.030
Text+Images 5+5	0.892	0.022

Factual Correctness (F1)

Model	Mean	Std Dev
ZeroShot	0.457	0.029
Text-Only	0.547	0.057
Text+Images 10+10	0.583	0.041
Text+Images 5+5	0.545	0.040

A.3. Storage Pipeline Diagram

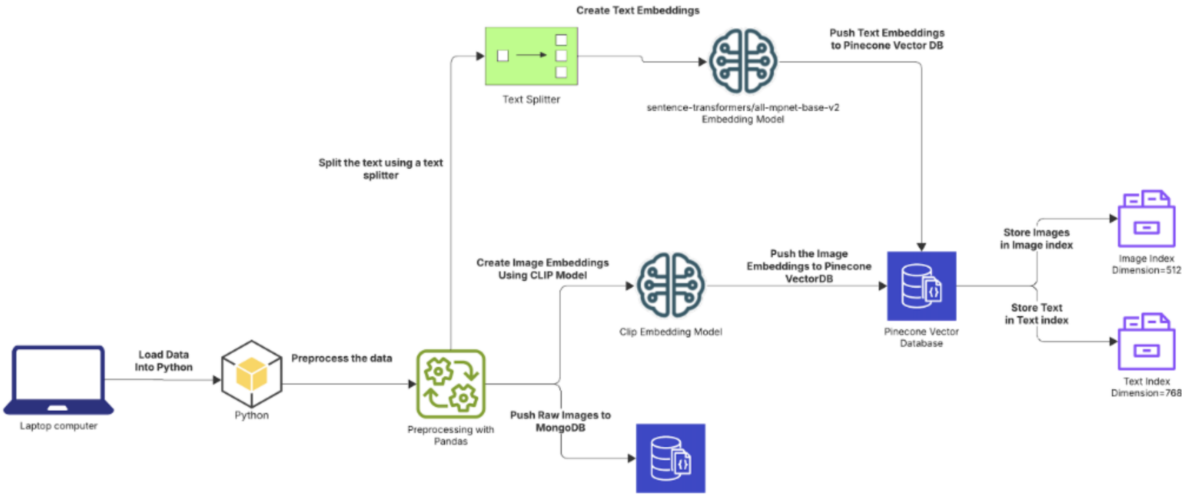
Create Text Embeddings
Push Text Embeddings to Pinecone Vector DB
Text Splitter
sentence-transformers/all-mpnet-base-v2 Embedding Model
Split the text using a text splitter
Store Images in Image index
Image Index Dimension=512
Create Image Embeddings Using CLIP Model
Push the Image Embeddings to Pinecone VectorDB
Clip Embedding Model
Pinecone Vector Database
Store Text in Text index
Text Index Dimension=768
Load Data Into Python
Preprocess the data
Laptop computer
Python
Preprocessing with Pandas
Push Raw Images to MongoDB

Figure 6: A.3 Pipeline of how we store our data

A.4. User Pipeline Diagram

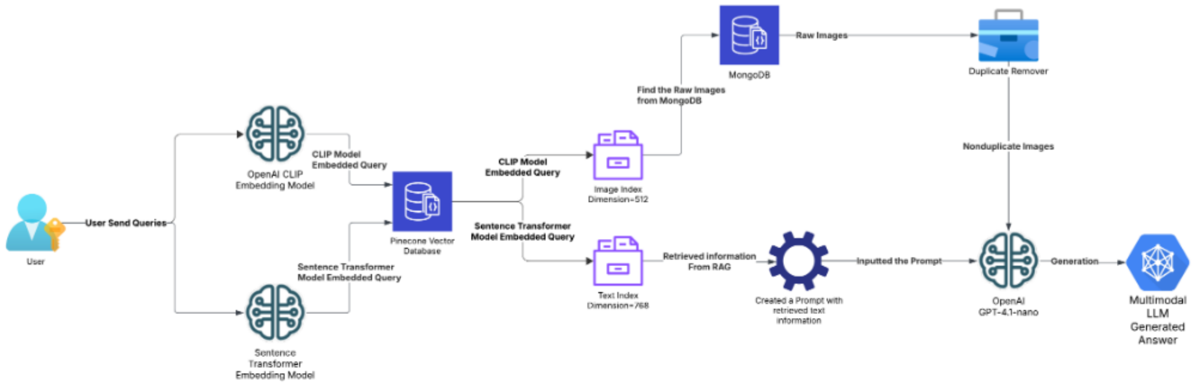
User
User Send Queries
OpenAI CLIP Embedding Model
CLIP Model Embedded Query
Sentence Transformer Embedding Model
Sentence Transformer Model Embedded Query
Pinecone Vector Database
CLIP Model Embedded Query
Sentence Transformer Model Embedded Query
Image Index Dimension=512
Text Index Dimension=768
Find the Raw Images from MongoDB
MongoDB
Raw Images
Duplicate Remover
Nonduplicate Images
Retrieved information From RAG
Created a Prompt with retrieved text information
Inputted the Prompt
OpenAI GPT-4.1-nano
Generation
Multimodal LLM Generated Answer

Figure 7: A.4 Pipeline of how the user is going to experience the architecture