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 an approach to training Large Language Models (LLMs), has emerged as a powerful framework to ground generation in trusted content. In the educational context, these are materials sourced by professors/teachers for a specific courses. Although RAG systems traditionally rely on textual input, modern digital textbooks often include 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 to build an intelligent assistant.

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 several versions of text only and multimodal Large Language Models and evaluation 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 questions but not for generic, 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

CEUR-WS.org/Vol-4010/itb25_s2s3.pdf

1. Introduction

Recent advances in Large Language Models (LLMs) 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 as seen in appendix 3 figure 6 and appendix 4 figure 7.

This is an exploratory study designed to position a deeper understanding on potential technical approaches for developing an LLM based Intelligent Assistant to support students with a focus on

iTextbooks'25: Sixth Workshop on Intelligent Textbooks, July 26, 2025, Palermo, Italy

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).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

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 an exploration on whether the incorporation of visual content enhances the generation of educational response and under what circumstances it may hinder it. The RAG-bot is specifically designed for an 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) [10]. The rise in the presence of AI in global society and its emergence in our daily 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 [8]. 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 needed. Ouyang and co-authors made note of this point in their construction of an AI literacy framework through a meta-analysis of papers spanning several disciplines [10]. 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 [13].

The nature of how AI driven tools are 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 courses, generating 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 [9]. The follow-on work will not only present the tool, but will give students an 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 an active learning approach known to be productive for learning in STEM environments [1].

The original RAG framework [7] 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 [11]. 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 (refer to Appendix A.3 and A.4 for model architectures).

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 originally accessed in PDF format. Additionally, audio recordings from lecture videos were transcribed into text using YouTube's speech-to-text transcription tool and incorporated as part of the textual dataset [14].

3.2. Embedding Details

Textual Data: As shown in Appendix A.3 figure 6, text chunks (1500 tokens, 100-token overlap) were embedded using SentenceTransformer `all-mpnet-base-v2` and stored in a text-only Pinecone Database (dim=768).

Visual Data: As shown in Appendix A.3 figure 6, to support multimodal retrieval in a RAG pipeline, we leverage OpenAI's CLIP model, a pretrained model which embeds texts and images into the same vector space and minimizes the distance between semantically similar image and text vectors [12]. CLIP has proven effective in zero-shot image classification, reducing or eliminating the need for expensive training on application-specific image datasets. This alignment enables cross-modal similarity comparisons: both textual passages and visual assets (e.g., images) can be encoded using CLIP's respective encoders and stored as embeddings in a vector database. At inference time, a user's natural language query is encoded via the text encoder and used to identify embeddings using nearest neighbors. This enables semantically consistent retrieval across modalities—e.g., matching a query like "building a decision tree" to visual representations of decision tree materials. Results are then passed to a language model for generation.

Retrieval Method: While querying to the RAG system, the prompt is embedded using both models (SentenceTransformers and CLIP). As illustrated in figure 7 in Appendix A.4, this creates a dual process that ends with feeding a multimodal LLM with supplemental content from both the image and text databases. The reason for keeping two separate databases is to allow for the comparison of text only versus text plus image generation. The raw images are stored in MongoDB for retrieval post the embedding search phase. The name of each image is stored as the metadata in the image database. Upon completing the 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. This treatment is the baseline for how the LLM responds to the query without additional content added to the question passed to the LLM.
- **Text Only RAG (10 text vectors)**: Using only text retrieved from our vector DB. This treatment includes the user's query along with the top 10 text vectors retrieved by highest cosine similarity score to the user's query.
- **Balanced Swap (5 text + 5 image vectors)**: Text with less relevant vectors replaced by top images. This treatment includes the user's query along with the top five text vectors retrieved by highest cosine similarity score to the user's query and the top five images retrieved by highest cosine similarity score to the user's query.
- **Text + Image (10 text + 10 image vectors)**: Addition of more visual information along with base textual information. This treatment includes the user's query along with the top ten text vectors retrieved by highest cosine similarity score to the user's query and the top ten images retrieved using nearest neighbors as the model and cosine similarity as the distance measure to the user's query.

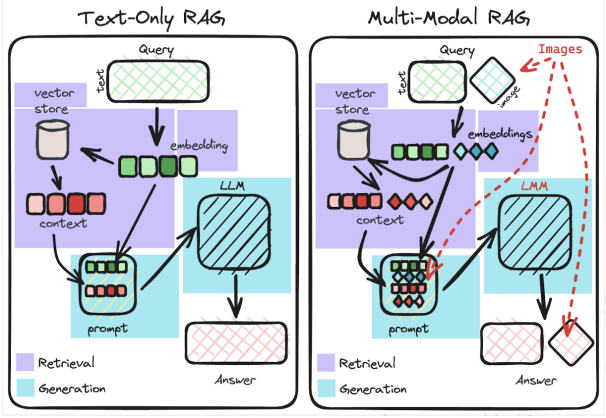

Figure 1: Visualization of how multi-modality differs from text-only RAG

The evaluation dataset was constructed to reflect course-aligned content, incorporating both generic and specific questions derived from educational materials. Generic questions are designed to be answerable using general knowledge, while specific questions require information from unique sources such as lecture notes, specific textbooks, and lecture recordings. Consequently, standard LLMs may under perform on specific questions compared to generic ones, highlighting the importance of RAG in generating accurate responses (See appendix A.3 and A.4).

To ensure the questions' specificity and relevance, we curated a set of 30 questions, 15 specific and 15 generic. This approach was preferred over mass-generation using LLMs, as LLM-generated questions may lack the desired specificity and could lead to inconsistent answers for evaluation. These questions and answers were generated using the authors' expertise as graduate students and then cross-reference through multiple School of Data Science faculty. The model-generated answers to the questions serve as the key metric for evaluation. Bootstrap resampling method allows for a robust measure of the quality of the responses. Additionally, we imposed word limits on the answers to further reduce variability in the system's responses, enabling a more controlled assessment of its performance.

3.4. Evaluation

Conext Recall, Faithfulness, and Factual Correctness were chosen to target three critical components of RAG for evaluation. We wanted to evaluate the key elements of the model's performance, seen in figure 2, in terms of the Context retrieved, the relevance to the Query, and the generated Response. The list below outlines how each metric aligns with one of these components. From the Context, we need to ensure that the generated response aligns with the retrieved information. From the Query, we need to ensure that the retrieved context is relevant to the query at hand. From the Response, we need to ensure that the response actually answers the query we began with.

We used the RAGAS [6] evaluation package to obtain metrics for our models:

- **Context Recall**: Measures whether the model successfully retrieved the right pieces of information. For example, when a user asks a question, the model pulls in background documents to help it answer. Context recall tells us what fraction of the relevant supporting documents were actually retrieved. In our project, this metric helped us assess whether models could correctly surface source material, particularly for domain-specific queries that rely on subtle or technical context (Context).
- **Faithfulness**: Evaluates whether the model's reasoning is grounded in the material it retrieved. Even if the model finds the right documents, it might still produce responses that are misleading or overconfident. Faithfulness measures whether the model's answer can be directly supported by the retrieved evidence. In our use case, we applied this metric to ensure that the generated answers didn't hallucinate facts or stray from the actual contents of the documents (Query).

- **Factual Correctness**: Assesses whether the final answer itself is accurate in relation to the query, even if the model uses correct information in the wrong way. This is the most outcome-focused of the three metrics: it checks if the model ultimately gives a factually valid response. For example, even if the right context was retrieved and used, the final output still needs to be judged on whether it answers the user's question truthfully. This was especially important for us when evaluating model responses to specific, high-stakes queries (Response).

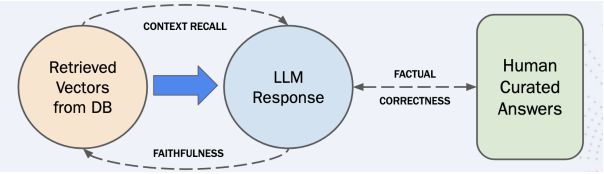

Figure 2: Visualization of how evaluation metrics interact with each component of the RAG model

In order to evaluate the three metrics, 30 questions were sampled 400 times for each of the four treatment groups, 200 for the generic and 200 for the specific. This resulted in a total of 1,600 scored responses: 800 from generic and 800 from specific questions. This number of model runs was chosen due to financial constraints associated with continuous use of the GPT API. We employed a pooled analysis strategy that aggregated all scores within each model and question type. This decision was guided by several factors: (1) all experimental runs were conducted under identical conditions with consistent question distributions and evaluation methods, (2) the goal of the study was to assess average model performance, and (3) the nature of LLMs allows for a certain amount of variability making for a robust estimate of model output distribution.

For each model and metric, we applied non-parametric bootstrapping by drawing 10,000 resamples with replacement from the scored responses to build an empirical distribution of the mean. From this, we computed 95% confidence intervals using the 2.5th and 97.5th percentiles of the resampled means. To compare models, we calculated the difference in their observed means and combined their bootstrap standard errors to form a confidence interval around the difference. If a model's bootstrapped confidence interval for the mean difference lay wholly above or below the baseline, it was judged significantly better or worse; otherwise, its performance was considered statistically indistinguishable from the baseline.

4. Experiment Results

We evaluated the effect of incorporating images into the RAG workflow 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 the curated dataset of generic and specific questions derived from course materials previously described. We compared these against a Text-Only RAG baseline and a Zero-Shot (no retrieval) setting. Evaluation was based on the three key metrics previously described:

Context Recall, *Faithfulness*, and *Factual Correctness*. Context Recall and Faithfulness are RAG specific measures and thus do not include the Zero Shot model.

In summary, the bootstrap-derived confidence intervals were narrow, and the statistical power of this design was more than sufficient to detect small-to-moderate differences in model behavior. This pooled analysis strategy is especially appropriate in studies like this one, where experimental conditions are controlled and average-case performance is the primary analytic focus.

4.2. Generic Questions

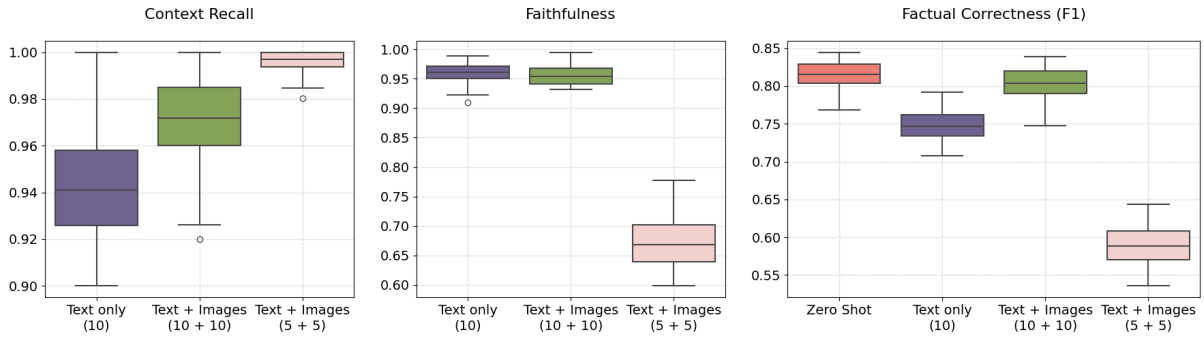

Figure 3: Performance Metrics on Generic Questions

On generic questions, Text-Only RAG performs competitively across all evaluation metrics, with no statistically significant differences observed when compared to either Zero Shot or Text + Images (10T + 10I) models for context recall, factual correctness (F1), or faithfulness. However, two statistically significant differences emerged. First, Balanced Swap (5T + 5I) performs significantly worse than Text-Only RAG on factual correctness (F1), with an average drop of approximately 16 percentage points ($p = 0.0008$, 95% CI: $[-0.25, -0.07]$). Second, the same model also underperformed significantly on faithfulness, showing a decrease of nearly 29 percentage points relative to Text-Only RAG ($p < 0.0001$, 95% CI: $[-0.42, -0.16]$). These findings indicate that while image augmentation at the Text + Images (10T + 10I) scale preserves parity with the baseline, the Balanced Swap (5T + 5I) configuration introduces meaningful degradations in factual accuracy and faithfulness.

For the RAG specific measures, we observed that the Text + Images (10T + 10I) configuration modestly improved Context Recall compared to the Text-Only RAG baseline. The Balanced Swap (5T + 5I) setup led to larger increase in recall, though not statistically significant, suggesting that more testing would be needed to validate that adding well-ranked images can improve retrieval relevance.

However, this improvement came at a cost. Faithfulness and Factual Correctness declined in the Balanced Swap (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 not significantly and 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 significant 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

Performance across models was generally similar to the Text-Only RAG baseline for both context recall and faithfulness, with no statistically significant differences observed. All models performed at or near

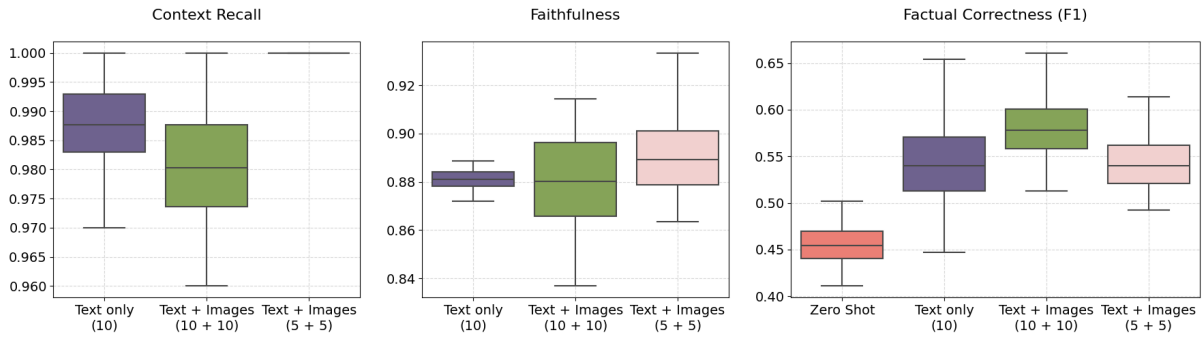

Figure 4: Performance Metrics on Specific Questions

ceiling for context recall, and faithfulness scores varied slightly but within overlapping confidence intervals. However, a significant improvement emerged for factual correctness (F1): the Text + Images (10T + 10I) model outperformed the Zero Shot baseline by approximately 13 percentage points ($p = 0.014$, 95% CI: [+0.03, +0.23]), indicating a meaningful benefit from richer context integration. As seen in the graphic above the CI between Zero Shot and Text + Images (10T + 10I) do not overlap. The Text-Only RAG and Balanced Swap (5T + 5I) models also showed improvements in factual correctness relative to Zero Shot, but these differences did not reach statistical significance. Overall, only the Text + Images (10T + 10I) configuration showed a robust advantage on specific factual accuracy.

This suggests that the inclusion of images under certain conditions in 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 in the Text + Images (10T + 10I) configuration slightly reduced context recall, likely due to visual noise introduced by less relevant images. However, the Balanced Swap (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.

When 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 certain quality measures of a multimodal LLM, especially for conceptually dense context. This is important when utilizing course materials such as lectures or digital textbooks for the creation of intelligent assistants, as the multimodal approach does seem to have advantages but in limited context. It is also important to note that while we did not experience any hallucinations in our experiment, this approach is not designed to prevent hallucinations from occurring, though the Faithfulness measure is design to quantify false or misleading content.

Specifically, our results also show that a fixed or naive strategy for image inclusion is suboptimal, meaning tuning the LLM to only include a limited and most relevant images is ideal. In the Text + Images (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.

6. Statement on Use of Generative AI

During the preparation of this manuscript, the author(s) used GPT4o to edit background context and summarize relevant research articles. The output was subsequently reviewed, revised, and fully controlled by the author(s). The authors take full responsibility for the accuracy and integrity of the content presented.

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] ExplodingGradients. *Ragas: Evaluation framework for LLM-generated responses*. https://docs.ragas.io/en/latest/. Accessed: 2025-07-08. 2025.

[7] 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).

[8] 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).

[9] 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).

[10] 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.

[11] 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.

[12] Alec Radford et al. "Learning Transferable Visual Models From Natural Language Supervision". In: *Proceedings of the 38th International Conference on Machine Learning*. 2021. URL: https://arxiv.org/abs/2103.00020.

[13] 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).

[14] Brian Wright. *MultiModalRAGbw*. https://github.com/NovaVolunteer/MultiModalRAGbw. Accessed: 2025-07-08. 2025.

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/

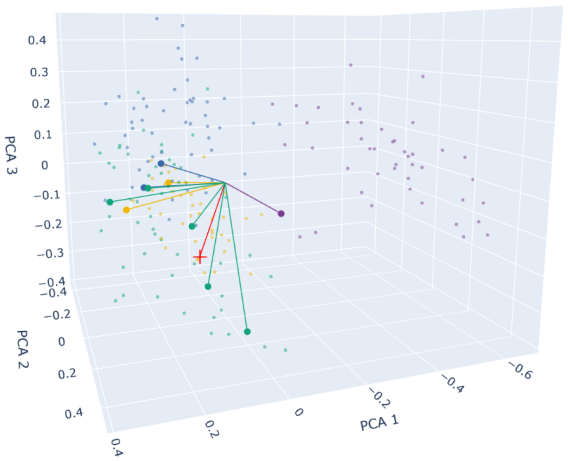

Figure 5: 3D Plot of PCA from 768 Dimension Text Vectors

A.2. Evaluation Metrics

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

A.3. Storage Pipeline Diagram

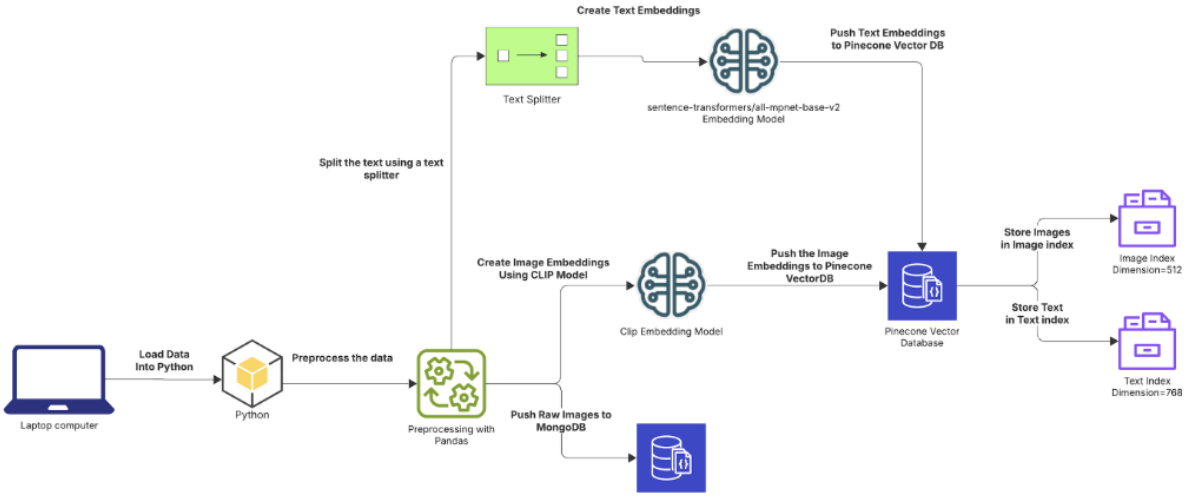

Figure 6: A.3 Pipeline of how we store our data

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.4. User Pipeline Diagram

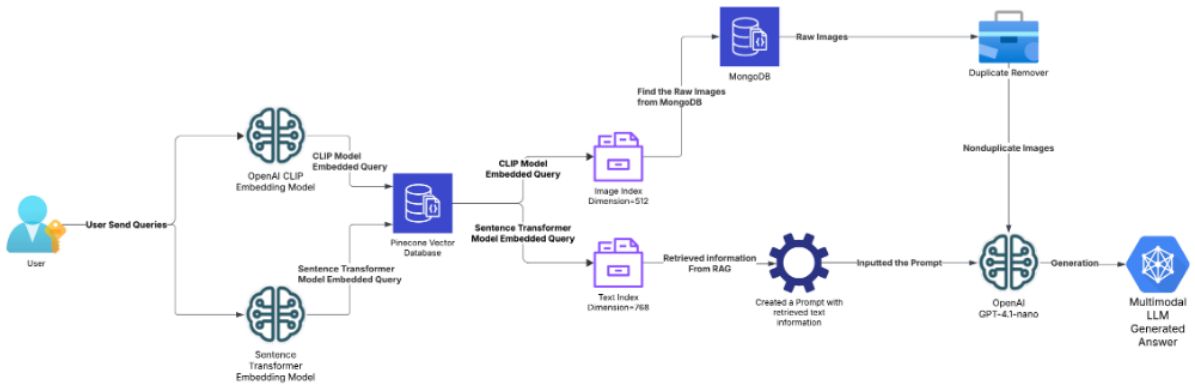

Figure 7: A.4 Pipeline of how the user is going to experience the architecture