UVA DATA SCIENCE

Using Educational Data to Explore Multimodal (Audio, Visual, & Textual) LLM Retrieval Techniques (to Enhance Textbook Utility)

Brian Wright, PhD
Quantitative Foundation Associate Professor
School of Data Science
University of Virginia

07/25/2025

Not Another ChatBot!

Trying to build a tool that is tailored to course content (textbooks, lectures, slides) and references it while the students are in the class – embed it to the digital textbook for the class

LLM?

- Are LLMs autoregressive?
- Can you control “how” autoregressivey they are?
- Do they often plateau (diminishing returns)?

Yes, Kinda of, Yes

Project Purpose

- Application of RAG for multimodal data retrieval with **Education Data (textbooks, slides and other digital content)**

- Is there **value in adding image content**?
  - **Measure the Quality** of responses, to help with the plateau

Assumption:

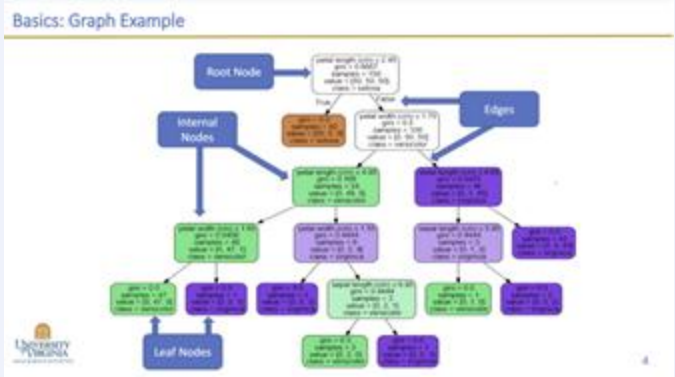
- This RAG-bot is scoped for an undergraduate ML course, nothing more or less advanced

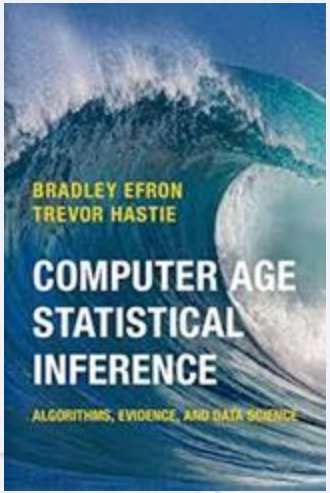
BLUF

- We tested on simple and specific questions
- Simple – questions don’t need images, but it doesn’t hurt the response
- Specific questions seem to…for now… perform better with the images

Data Discussion

- Course content of **DS3001** (Foundations of Machine Learning)
  - Slides (Text and Images)
  - Lectures (Audio converted to Text)
  - ML research papers/Data Science textbooks (Text and Images)





Embedding Details

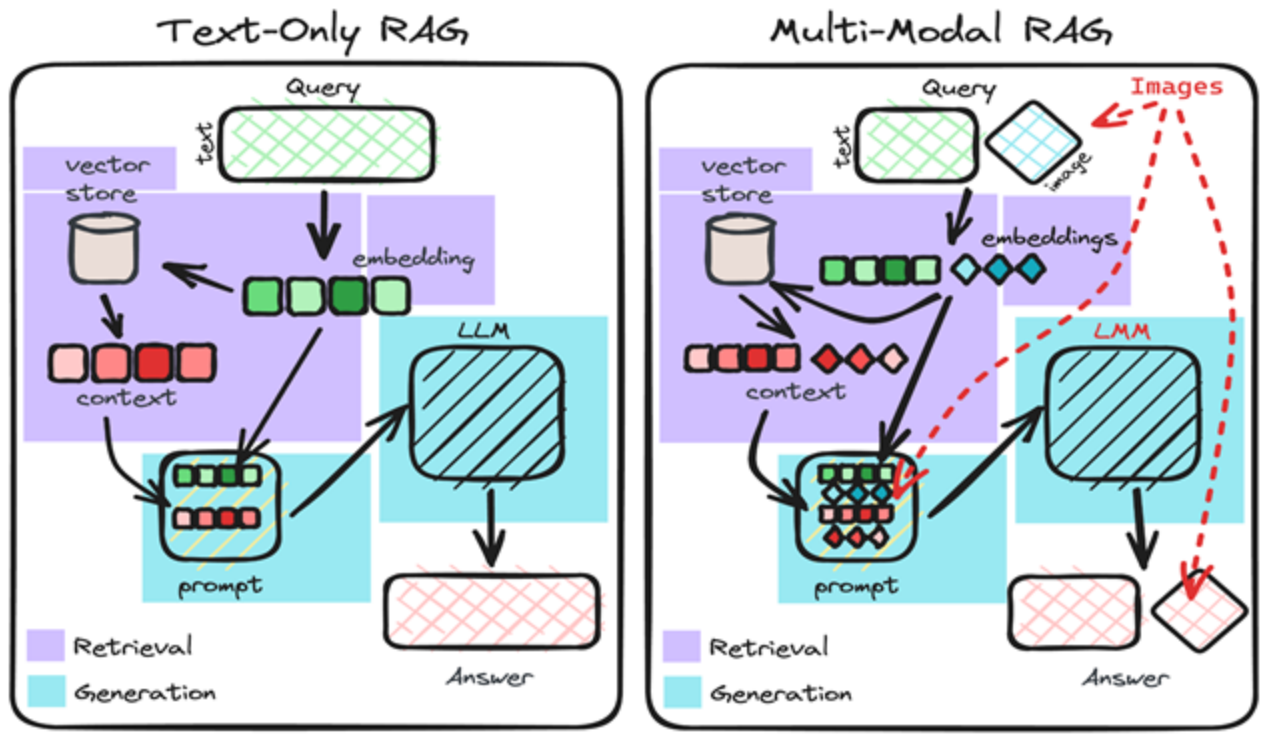
TEXT DATA

- Embeddings into **Pinecone DB**
- Model: SentencesTransformer *(all-mpnet-base-v2)* - Dim:768
- Text Chunks Size: 1500 Tokens
- Text Overlap Size: 100 Tokens
- Number of Embeddings: 5096

IMAGE DATA

- Embeddings in **Pinecone DB**
- Model: OpenAI Clip Model *(clip-vit-base-patch32)* - Dim:512
- Actual Images in **MongoDB** Retrievable based on metadata
- Number of Embeddings: 1169

RAG Workflows



Designing Our Experiment

- **Zero Shot – No training**
- **Text Only RAG**
  - 10 Vectors (word chunkcs
- **Text + Image RAG**
  - **Text + Image (10T + 10I):** Adds 10 image vectors to the 10 retrieved text vectors, preserving all textual context while layering on visual information.

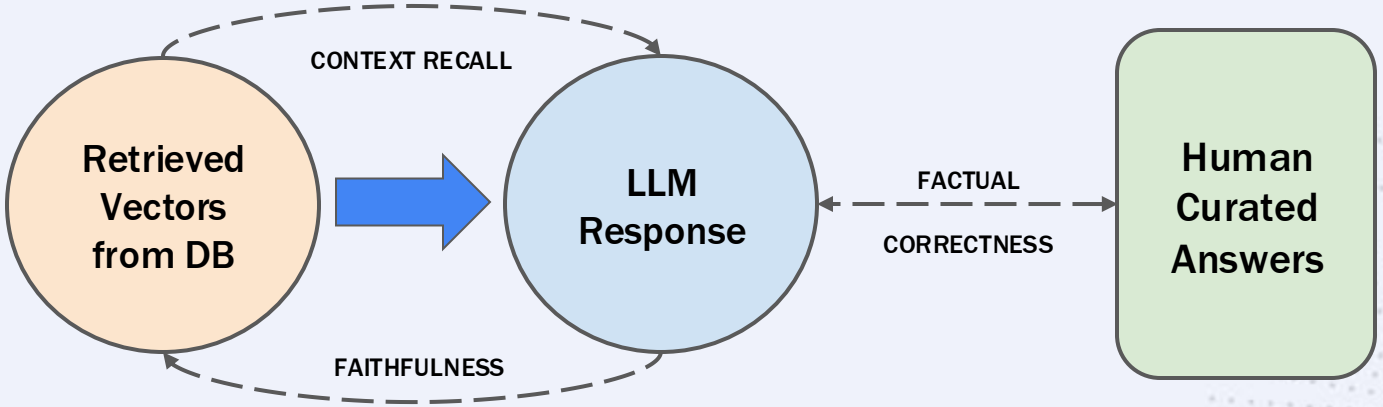
  - **Balanced Swap (5T + 5I):** Replaces the bottom 5 text vectors with the top 5 image vectors, maintaining the same number of total context inputs but altering the text-image ratio.

RAGAS (Evaluation Package: https://docs.ragas.io/en/stable/

Evaluation Phase

Compared model performance across 3 metrics

- **Context Recall** - how many of the relevant documents were successfully retrieved.
- **Faithfulness** - are all claims in response supported by the retrieved context.
- **Factual Correctness** - factual accuracy of the generated response with the reference



https://docs.ragas.io/en/stable/concepts/metrics/available_metrics

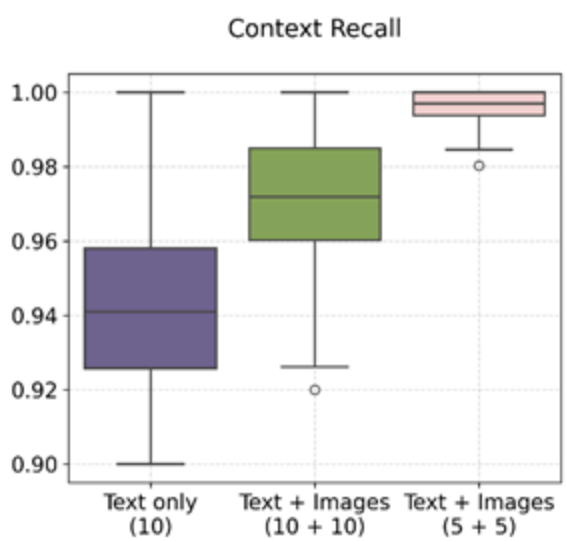
Evaluation Dataset

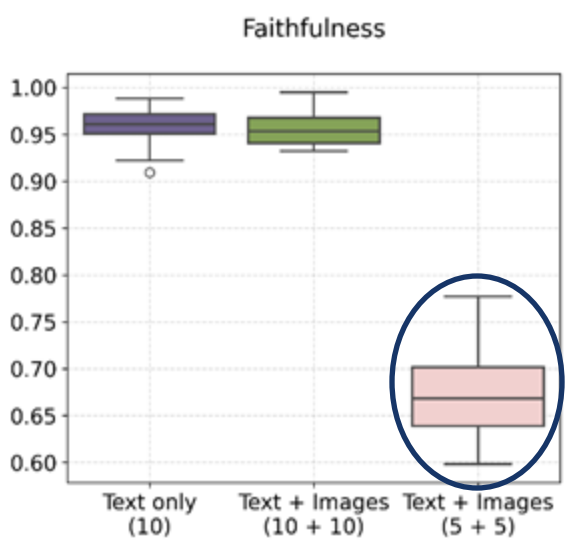
- Combination of **general data science** and **highly domain-specific** question.
- 30 questions total, 15 specific and 15 generic
- Conducted **bootstrapped testing** – 30 questions were sampled 400 times for each of the four treatment groups, 200 for the generic and 200 for the specific. This resulted in a total of **1,600 scored responses**: 800 from generic and 800 from specific questions

Results & Discussion

RAG Performance on Generic Questions

*Bootstrapped RAGAS using GPT 4.1-nano
50 questions - 10 iterations





➔ Increases on addition of images

➔ Further increase when less relevant text is replaced with highly relevant images

Images are useful for LLM retrieval

➔ Marginally different on adding images

➔ Big drop when bottom 5 Text vectors replaced with images

The Bottom 5 Texts are more important for the response

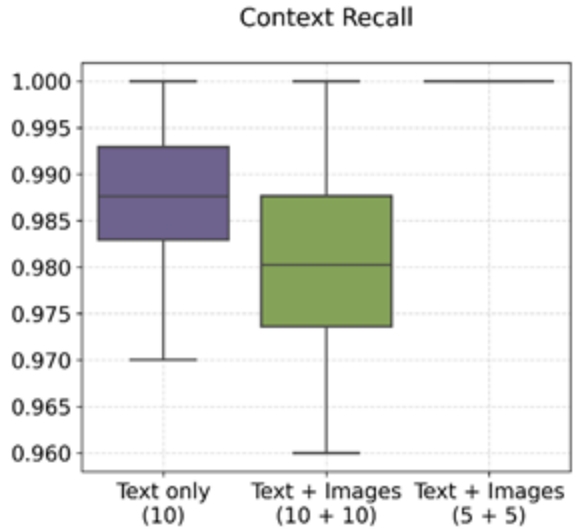
➔ Zero-shot is pretty good because questions are generic ML 101

➔ Increases on addition of images
Images help add context to text-only response

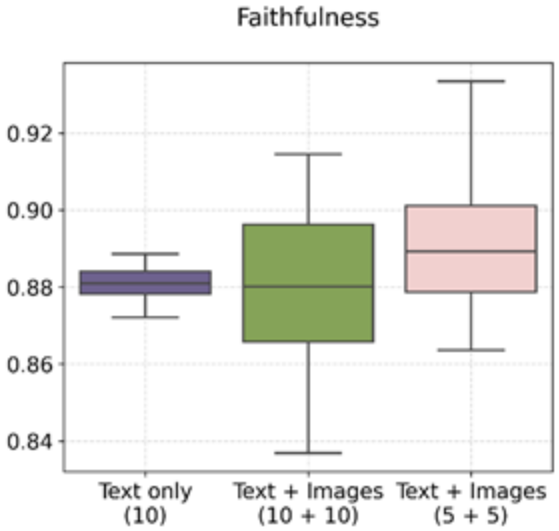
➔ Drops when bottom 5 text vectors are replaced with images

*Bootstrapped RAGAS using GPT 4.1-nano
50 questions - 10 iterations

RAG Performance on Specific Questions



- ➔ Recall decreases on addition of images
  Too many images slightly confuses the LLM
- ➔ Perfect Recall when only top text and images provided



- ➔ Adding images improves adherence to retrieved contextual information
- ➔ Slightly increases when bottom 5 Text vectors replaced with images

Specific Questions need lesser scoped contexts

- ➔ Zero-shot is comparatively bad at specific questions *as expected*
- ➔ Increases on addition of images
  Images help add info to response

Discussion

There is varied performance based on the ***specific*** queries - helps with the plateau

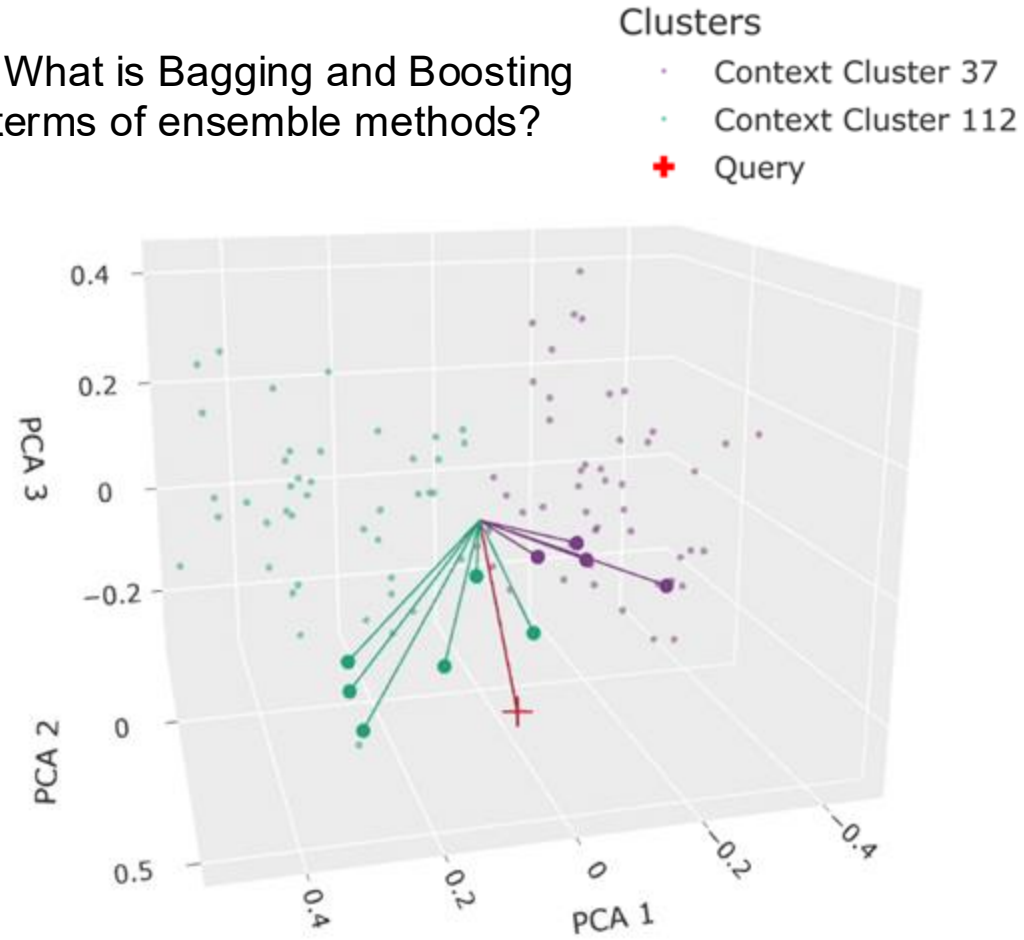
GENERIC QUERIES	SPECIFIC QUERIES
Adding images to current text *doesn't hurt* the LLM	Adding too many images *confuses* the LLM
Replacing less relevant text with images *hurts* the performance	Replacing less relevant text with images *boosts* the performance

Modeling

1. Testing the RAG Setup

- Using Manual Prompting
- Embedding Visualization
  - Concepts Clustering (HDBScan - KMeans)
  - PCA to bring down dimensions from 768 to 3
  - Note the 10 dots

Q: What is Bagging and Boosting in terms of ensemble methods?

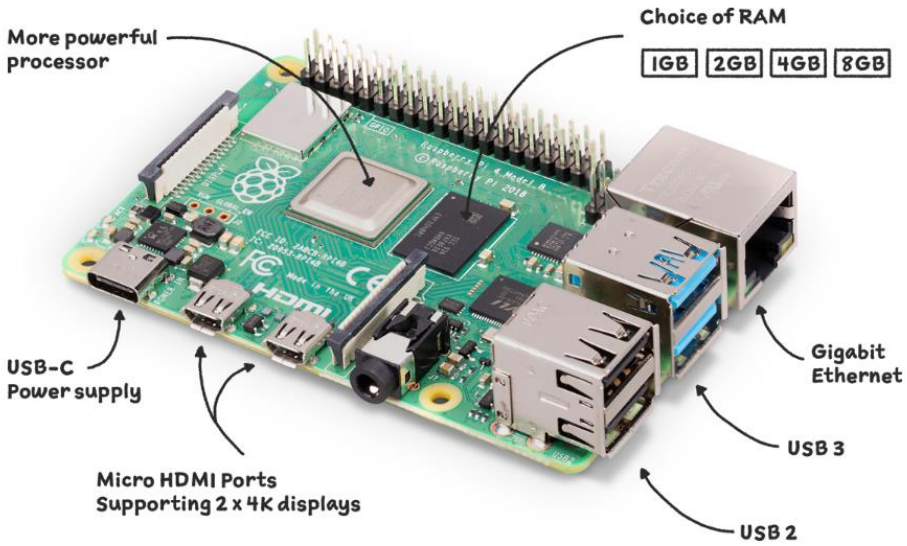


Future Work and Things to Consider

Future Work in this project

- **Working to develop for use in the classroom**
  - Focus on student self regulated learning
  - Design - 2 courses one with the tool or without

- **Develop hardware to be placed in the class during lab periods**
  - Use rasberrypi's will pretrained llm fron the course content



Future Work

Thank You! Questions?

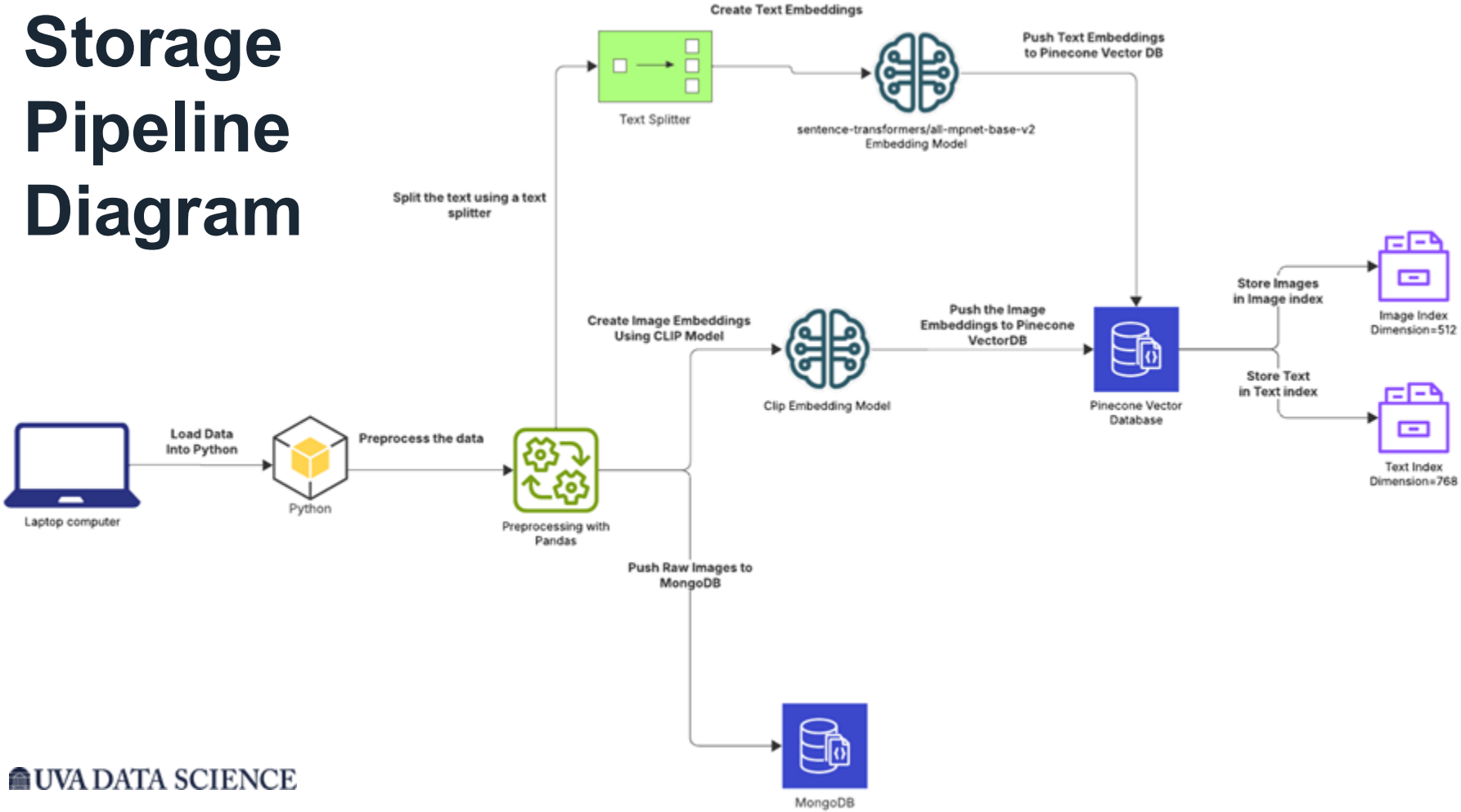



Vishwanath Guruvayur Luke Napolitano Doruk Ozar Bereket Tafesse

The Research Team, MSDS Students

UVA DATA SCIENCE

Storage Pipeline Diagram



User Pipeline Diagram