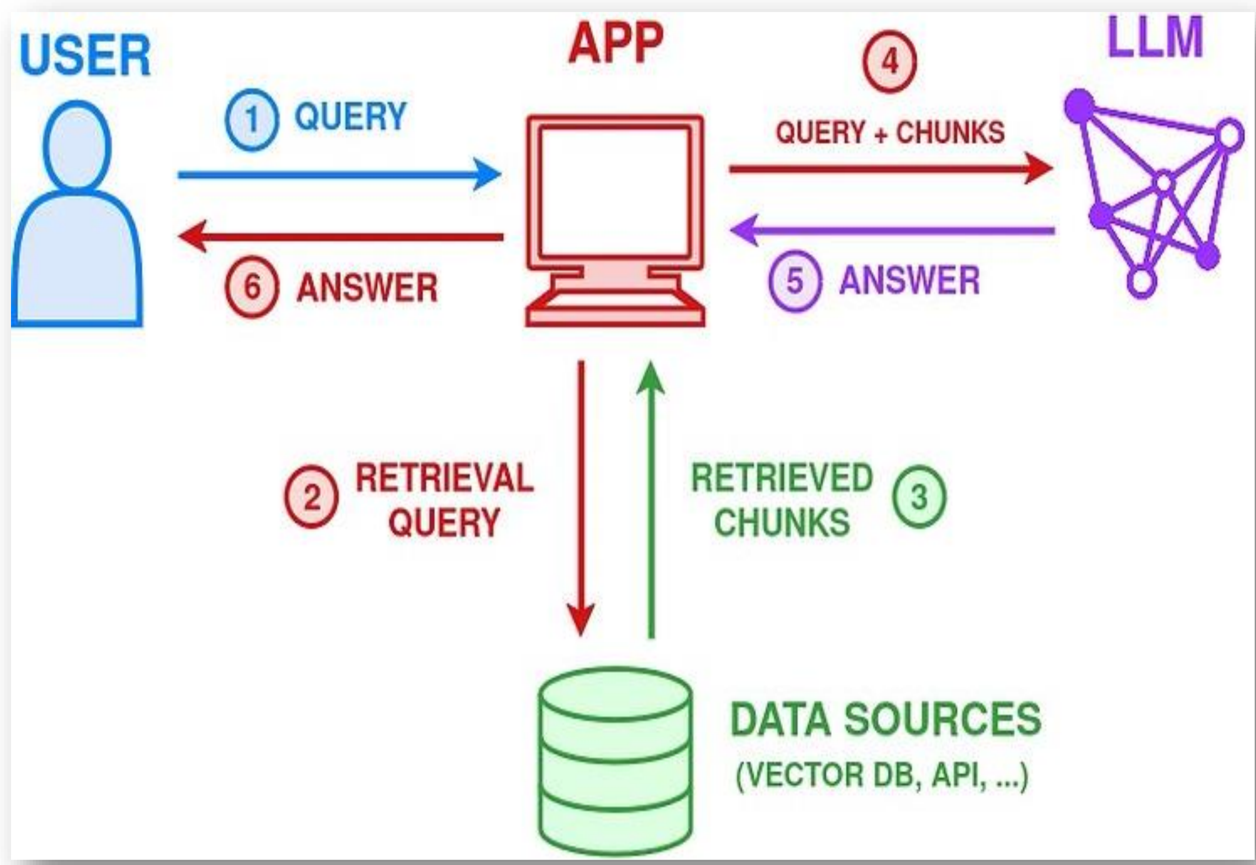
Development of an Artificial Intelligence RAG Chatbot for Use as a Surgical Reference Tool

McKenna Parker (SEAS 2027), Kevin Du (SEAS 2027), Daniel Weber (BE), Kristoffel Dumon (MD), Noel Williams (MD)
Department of Surgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA

Introduction and Background

Artificial intelligence (AI) and large language models (LLMs) have become increasingly viable as a method of learning and studying material across all topics. One way to integrate AI is through retrieval augmented generation (RAG), which allows developers to input data that the LLM should use as a reference when answering an end user's prompt. This study aimed to develop a RAG-enabled AI chatbot using surgeon-approved documents on airway trauma surgeries.

Figure 1: Architecture of a basic RAG system. *Retrieved from ridgerun.ai.*[1]



Methods

Python
- Programming language used to write code base

LangChain
- Open-source framework to aid in development of applications with AI integration

OpenAI API
- LLM used to create embeddings and answer prompts

Pinecone
- Vector database for storing embeddings and doing similarity searches

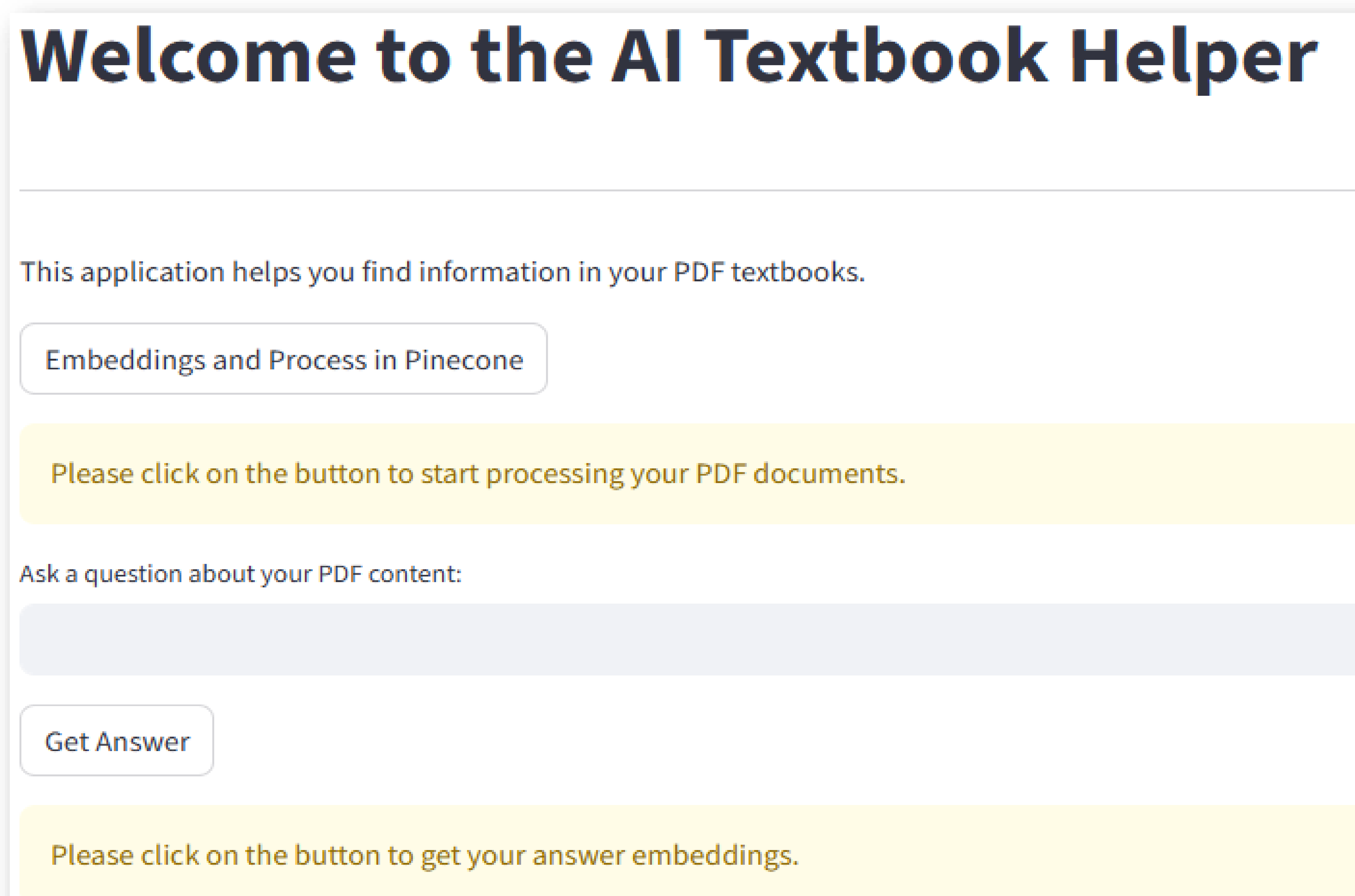
Streamlit
- Python web app hosting for deployment

Figure 2: RAG-enabled chatbot response to the prompt "What is a tracheostomy?"

A tracheostomy is a surgical procedure that involves creating an opening in the front of the neck to access the trachea (windpipe). This procedure is indicated for patients who require prolonged intubation or have conditions that obstruct the upper airway, such as tumors, severe infections, or trauma. It involves making an incision in the neck, dissecting through layers of tissue to reach the trachea, and inserting a tracheostomy tube to secure an airway.

The information provided is contained in the document titled "OPEN MANUAL OF SURGERY IN RESOURCE-LIMITED SETTINGS" by Richard Davis and Joseph Nderitu, found under the section "Introduction" and "Tracheostomy proceeds in the following general steps:" of the manual.

Figure 3: User interface of the Streamlit application that hosts the AI Textbook Helper.



Results

After uploading three PDFs on Cricothyroidotomy[2], Tracheostomy[3], and Airway Management in Trauma[4], the RAG-enabled chatbot was able to return viable information regarding these operations. Using built-in prompt editing, reference information was also provided with each answer to a user prompt on relevant information.

Next Steps

This project was the first iteration of a minimum viable product in a larger framework. In the future, the lab hopes to build from findings to develop a desktop application with AI RAG integration that searches a vast database of approved resources including but not limited to videos, anatomy models, textbooks. The first step in this process is creating a similar project within the Unity development environment using Pinecone's .NET software development kit. Another consideration for future iterations is accuracy and pricing of different AI models.

Perelman School of Medicine, University of Pennsylvania

PURM Penn Undergraduate Research Mentoring, Center for Undergraduate Research & Fellowships

References

[1] Araya, A. (2024, November 14). *How to Evaluate Retrieval Augmented Generation (RAG) Systems*. RidgeRun.ai.
[2] Davis, R. (2023, October). *Cricothyroidotomy*. Vanderbilt University Medical Center.
[3] Davis, R., & Nderitu, J. (n.d.). *Tracheostomy*. Vanderbilt University Medical Center.
[4] Sund, G., Kynes, M., & Yao, P. (2022, November). *Airway Management in Trauma*. Vanderbilt University Medical Center.