RadPointGPT: A Generative Chatbot to Facilitate Access to Department SOPs

Paulo Kuriki, MD, Neuroradiology Fellow, Radiology, UTSW
Fernando Kay, MD; Cecelia Brewington, MD; Ronald Peshock, MD

Background/Problem Being Solved

Standard Operating Procedures (SOPs) are crucial for establishing guidelines and protocols in radiology. However, they often result in an overwhelming number of documents, making it challenging for staff to locate specific information. Traditional search methods struggle with natural language queries especially amidst vast amounts of data, highlighting the need for a more effective retrieval system.

Intervention(s)

We developed a chatbot that combines Large Language Models (LLMs) with the Retrieval Augmented Generation (RAG) method to improve access to SOP documents. We converted SOPs to plain text, tokenized them, and stored them in a vector database, enabling similarity search for data retrieval. A web chatbot interface allows users to interact with documents using natural language. Questions are queried to the vector database, which retrieves context chunks, creating a prompt that is sent to a locally deployed Mistral-7B model, generating an answer for the user. This approach was chosen for its speed, performance, and security in handling sensitive information.

Barriers/Challenges

The RAG method can struggle with a large number of documents, such as the 252 SOP files in our database. As the document count increased, accurately selecting the appropriate document to provide the necessary context for the model proved challenging.
We implemented an Evaluation Process using GPT-4 to assess document retrieval and response accuracy. We generated 10-20 questions per document and submitted them to the RAG+LLM pipeline, where each response was evaluated for retrieval accuracy and response relevance.

Outcome

The Evaluation Process allowed us to identify and quantify failures, leading to significant enhancements, including embedding optimization, question refactoring, retrieval re-ranking, metadata improvements, and SOP error corrections. As a result, the global F1-Score reached 93.5% with an accuracy of 87.8%.

Conclusion/Statement of Impact/Lessons Learned

Integrating the RAG process has revolutionized document retrieval in our department. The locally deployed Mistral LLM enables easy access to SOPs while ensuring data confidentiality. The Evaluation Pipeline was crucial for identifying errors and enhancing our system's efficiency. Future plans include refining the retrieval process, addressing biases, and incorporating user feedback for continuous improvement.

Synopsis

This work describes the development, validation, and deployment of a system that integrates a chatbot powered by LLM, and RAG to enhance access to SOPs in a radiology department.
Featuring a chatbot and a locally deployed Mistral-7B model, it effectively addresses large-volume document retrieval challenges and has shown significant improvements after an Evaluation Process was implemented, achieving a global F1-Score of 93.5% and an accuracy of 87.8%.

Figure(s)

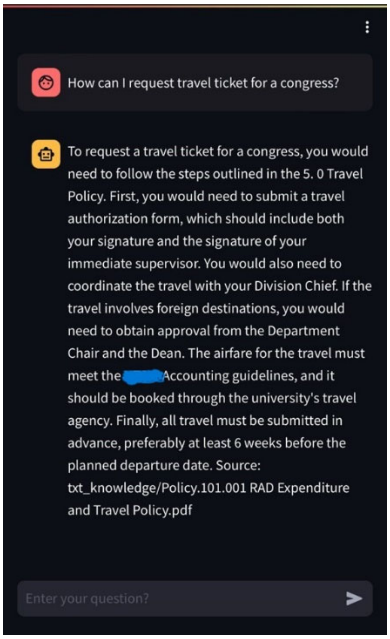

Figure 1. The screenshot shows the chatbot interface where users can request SOP information. System transparency is maintained by linking to the original document source. Attendees will be provided with a link to test the classifier during the presentation.

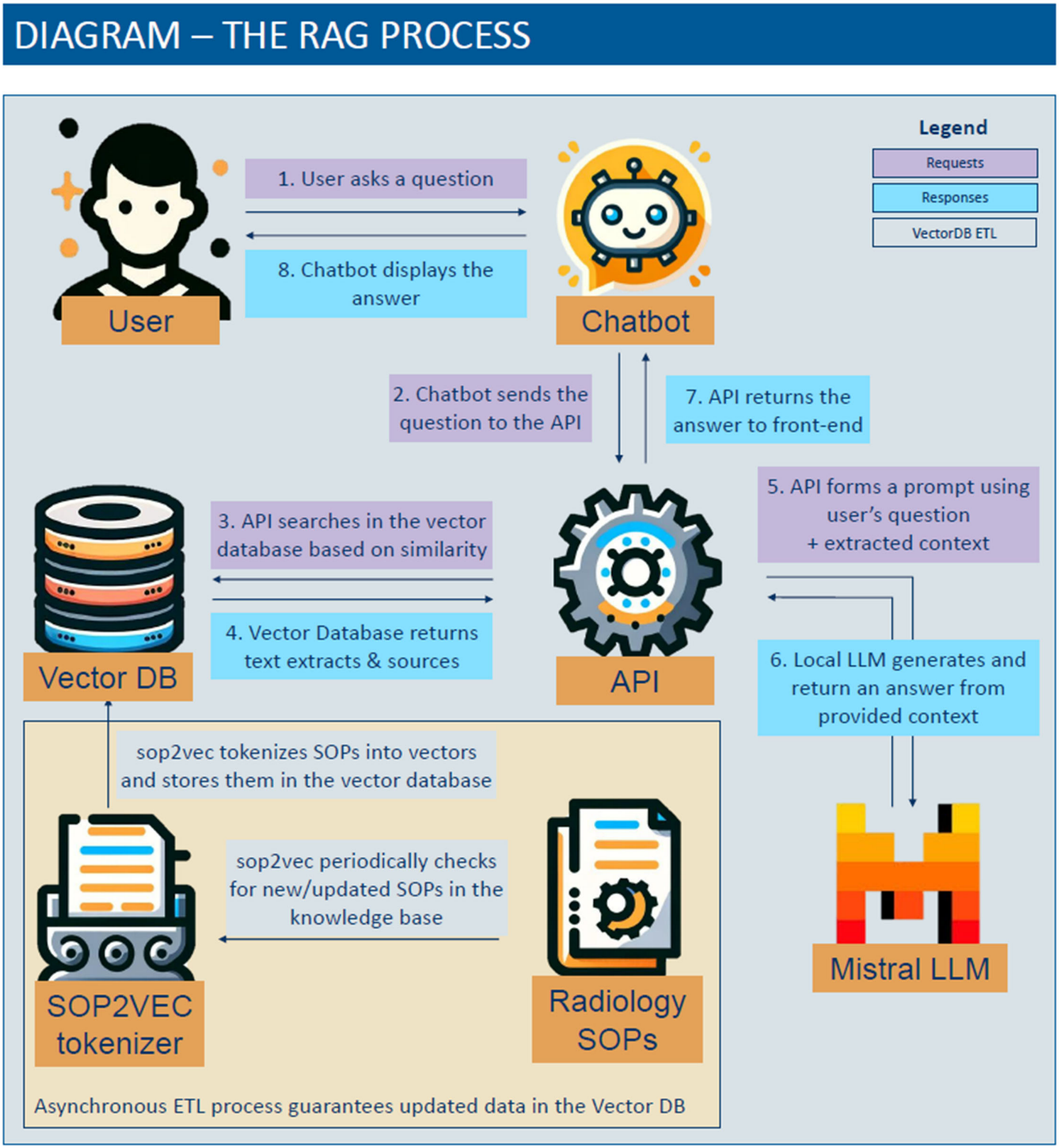

Figure 2. This diagram illustrates the flow of requests and responses between the user, chatbot, API, vector database, tokenizer, SOPs repository, and the local LLM within the Retrieval Augmented Generation (RAG) process.

Keywords

Administration & Operations; Applications; Artificial Intelligence / Machine Learning; Clinical Workflow & Productivity; Communication Data Management; Educational Systems; Emerging Technologies; Organizational & Professional Development; Quality Improvem