SIIM CAIMI26 Conference on Artificial Intelligence in Medical Imaging
OCT 26-27 | University of Pennsylvania, Philadelphia, PA

Retrieval Augmented Generation Enables an AI-Powered Radiology Institutional Knowledge Assistant

Amol Sinha, MS, Stanford Medicine Children's Health; Liliana Ma, MD, PhD; Sergios Gatidis, MD; Sivashankar Dhandapani; Kenneth Wong; Shreyas Vasanawala, MD, PhD

Introduction/Background

Retrieval Augmented Generation (RAG) mitigates LLM hallucinations by grounding models with reliable context to improve accuracy. Our Radiology department relies on an organically growing, SharePoint site containing >15 years of institutional knowledge, including 6.23 Gb of inconsistently structured data (942 webpages and >5000 documents). However, this site's scale has grown beyond the capabilities of standard indexing tools. Thus, the goal of this study was to pilot a RAG-based chatbot to centralize and streamline access to this repository, enabling natural language question-answering for clinical staff and trainees.

Methods/Intervention

This project was performed in the pediatric radiology department of an academic hospital following a pilot PDSA (Plan-Do-Study-Act) cycle. A RAG-powered chatbot was deployed within an enterprise Azure, HIPAA-compliant server (Figure 1). Institutional SharePoint documentation was extracted to Azure blob storage. User queries were compared with the SharePoint database to retrieve the eight most relevant document chunks. These chunks, along with the initial query, were fed to the LLM, which generated summarized answers containing citations and direct hyperlinks to the source pages. A 25-question validation set was sourced from trainees and attendings, covering clinical protocols (n=11), institutional policies (n=9) and clinical procedures (n=5). One diagnostic radiology resident trainee (PGY-4) and two attending reviewers, with 15-25 years of imaging experience, graded the responses from the RAG pipeline using a 5-point Likert Scale deriving a “User Effort Score” (UES; Figure 2). Inter-rater reliability among the three reviewers was calculated using a two-way random-effects model for absolute agreement.

Results/Outcome

The tool achieved an aggregate UES of 3.97±1.30 (range: 3.76-4.16, Figure 3). 68% of all responses were rated as UES≥4. Inter-rater reliability demonstrated moderate agreement among reviewers (ICC=0.68, CI: 0.48-0.83). Based on these results, a feedback form was added to the chatbot for further iteration and refinement.

Conclusion

Initial results suggest robust utility across training levels. Future iterations will focus on quantifying institutional time-savings versus manual SharePoint navigation.

Statement of Impact

This pilot PDSA cycle demonstrated that a secure, RAG-based chatbot can effectively navigate complex

institutional databases that are difficult to search using conventional indexing tools. This approach is generalizable to any institution managing large-scale clinical knowledge bases.

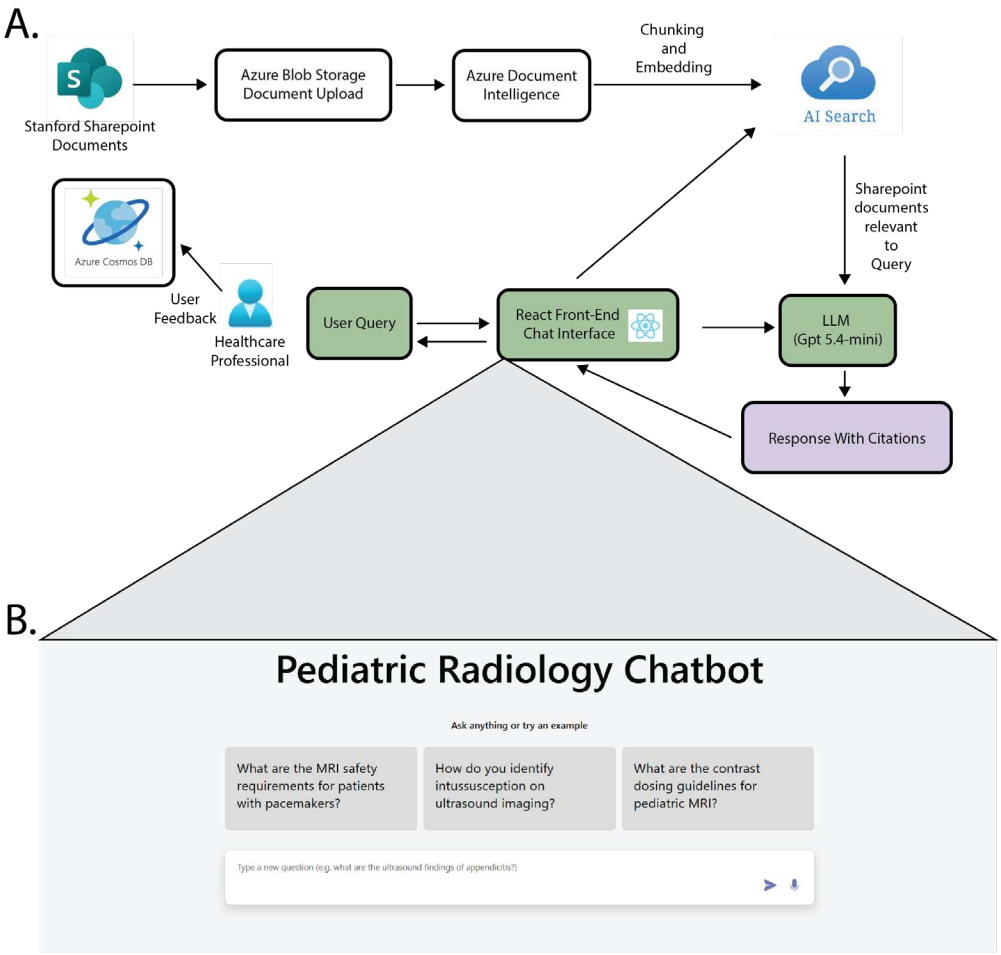

Figure 1: A, overview of RAG-based chatbot implementation within an enterprise Azure, HIPAA-compliant server. Azure AI Search was used for both embedding and retrieval, pulling relevant information from the Azure Blob storage. Context relevant to each user query was retrieved using Azure AI Search fed to the LLM for generation of a “grounded” response containing linked references. B, the interactive chatbot interface, created using the React framework (with institution-identifying information cropped).

Score	Impact on Workflow
5	**Extremely helpful:** Chatbot gave a great answer + the appropriate link. I didn't need to do anything else.
4	**Very helpful:** Chatbot gave a reasonable answer and appropriate link. I had to read the link for a few more details.
3	**Moderately helpful:** Chatbot probably saved me time; the Sharepoint link was relevant but I still had to click through related links get the full answer.
2	**Not helpful:** Chatbot gave a vague answer; I had to go to SharePoint and start a new search.
1	**Failed:** Chatbot was wrong or couldn't find a relevant document.

Figure 2: User Effort Scale details

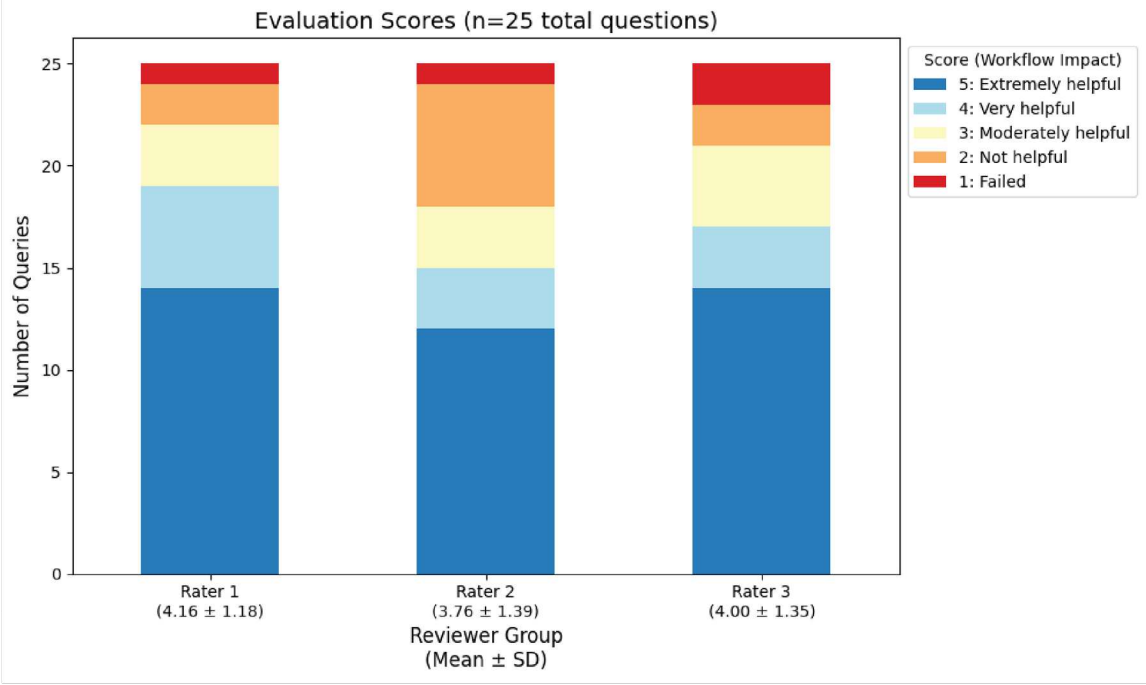

Figure 3: Breakdown of scores by each rater.

Keywords

Retrieval Augmented Generation; Large Language Models; Clinical Deployment; Protocol Optimization; Plan-Do-Study-Act