Agent-MedRAG: Design and Evaluation of an Agent-Driven Retrieval-Augmented Question Answering System Based on Biomedical Literature

Xinglan Zhao & Chuyang Su
Columbia University
New York, NY, USA
cs4570@columbia.eduxz3420@columbia.edu cs4570@columbia.edu
GitHub Link: https://github.com/skylarzhao1/GR5293-AgentMedRag

***Abstract*—As shown in recent work [?], ... This paper introduces a domain-specific Retrieval-Augmented Generation (RAG) agent designed for the biomedical field. The system addresses challenges in parsing, synthesizing, and reasoning over complex medical literature, leveraging recent advancements in large language models and structured retrieval pipelines.**

I. Motivation and Objectives

The medical domain is characterized by an overwhelming volume of highly specialized literature, including but not limited to clinical trial reports, epidemiological studies, basic biomedical research, and pharmacological data. Navigating this vast and complex information landscape poses significant challenges to healthcare professionals, who must rapidly interpret and apply critical knowledge within constrained timeframes. This bottleneck in information digestion can directly affect clinical decision-making and medical research productivity.

Recent advances in large language models (LLMs) and Retrieval-Augmented Generation (RAG) architectures provide a timely opportunity to address this challenge. In particular, tools such as LangChain enable the development of domain-adaptive agents that are capable of understanding, retrieving, and reasoning over biomedical literature in a structured and scalable manner. These systems have the potential to simulate the behavior of expert literature analysts, while offering the computational efficiency of modern AI systems.

To this end, the primary goal of our project is to design and implement a domain-specific RAG agent tailored to the biomedical domain. The system will be capable of parsing and synthesizing multiple types of medical documents—including clinical studies, case reports, and meta-analyses—while ensuring the interpretability and factual correctness of its outputs. In addition, the system will incorporate quality control mechanisms to rank and filter evidence, thereby promoting reliability and clinical relevance in retrieved responses.

TABLE I
Retrieval metrics on 4-question sample

Question	Precision	Recall	Most Relevant Document Summary (truncated)
Q1	1.00	1.00	Sonodynamic therapy (SDT) has emerged as a cutting-edge strategy for combating multidrug-resistant bacteria.
Q2	1.00	1.00	Calprotectin orthologs/paralogs have been identified in zebrafish ...
Q3	1.00	1.00	Vestibular migraine (VM) is a leading cause of recurrent vertigo episodes ...
Q4	1.00	1.00	Ni clusters act as transition materials between bulk and individual atoms, enhancing HER ...

II. Results and Interpretation

A. Retrieval Results

a) Interpretation.: All four queries achieved **perfect context-precision and context-recall (1.00)**, indicating that the agent retrieved *only* the passages required to answer—without missing any evidence nor introducing spurious text. This suggests our 7B-parameter model, together with a 2400-document corpus, is sufficient to guarantee exhaustive yet noise-free retrieval in this setting.

B. Generation Results

a) Interpretation.: While answer-relevancy remains high for three out of four queries, faithfulness notably degrades on Q3–Q4. Manual inspection shows numeric values were paraphrased incorrectly in these answers, even though the correct evidence was present—highlighting generation-layer hallucination.

b) Implications.: Perfect retrieval ensures clinicians are not burdened by irrelevant literature, but generation quality still limits direct clinical adoption. We plan to add

TABLE II
GENERATION METRICS ON 4-QUESTION SAMPLE

Question	**Faithfulness**	**Answer Relevancy**	**Context Summary (truncated)**
Q1	0.89	1.00	SDT has emerged as a cutting-edge strategy ...
Q2	0.88	0.93	Calprotectin heterodimer S100A8/S100A9 ...
Q3	0.57	0.00	Vestibular migraine (VM) as leading cause ...
Q4	0.54	0.96	Ni clusters improve hydrogen evolution reaction ...

citation-aware decoding and numeric consistency checks to mitigate remaining hallucinations (see Future Work).