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

Xinglan Zhao* Chuyang Su†
Graduate School of Arts and Sciences, Columbia University, New York, NY, USA
*xz3420@columbia.edu †cs4570@columbia.edu

***Abstract*—The rapid growth of biomedical literature makes it difficult for clinicians and researchers to locate and synthesise relevant evidence in a timely manner. We present Agent-MedRAG, a lightweight Retrieval–Augmented Generation (RAG) agent that reads PubMed abstracts, retrieves citable passages, and produces evidence-grounded answers to biomedical questions. A corpus of 2 400 PubMed articles (Jan. 2024–Jan. 2025) is split into 200-token chunks with 64-token overlap, embedded by `BGE-large-en-v1.5`, and indexed in a local Chroma vector store. Queries are semantically retrieved (Top–*k*) and re-ranked with a cross-encoder (`bge-reranker-base`); the top four passages are supplied to `Mistral-7B-Instruct` under a prompt that forbids external knowledge.**

Evaluation combines retrieval metrics—*Context Precision* and *Context Recall*—with generation metrics—*Answer Relevancy* and *Faithfulness.* On four representative biomedical queries the agent attains precision = recall = 1.00, demonstrating exhaustive yet noise-free retrieval. Relevancy remains high (≥ 0.93 on three queries), whereas Faithfulness exposes residual hallucination (0.57/0.54 on two complex questions), identifying generation as the current bottleneck. Qualitative analysis shows strengths in mechanistic reasoning and negation handling but highlights failures when reranking surfaces less-salient context.

Future work targets (i) scaling to larger backbones (e.g. `LLaMA-3-70B`) for deeper clinical reasoning; (ii) adding conversational memory for multi-turn refinement; (iii) implementing incremental indexing for real-time literature updates; and (iv) integrating evidence-grading frameworks such as GRADE to rank outputs by confidence level. These extensions aim to transform Agent-MedRAG into a continuously updated, trustworthy assistant for evidence-based medicine.

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.

II. Methodology

A. Data

We adopt **PubMed**[1], a public repository of biomedical literature, as our data source. Using a Python script, we randomly sampled 2 400 articles published between **Jan. 2024 and Jan. 2025**. For each article we extracted the *title*, *abstract*, and study *start/end dates.* This corpus is used to train an agent that reads abstracts and extracts key medical information. We employ `Mistral-7B-Instruct`[2] as the primary generator.

B. Model Architecture and Implementation

Our system follows a domain-adapted *Retrieval-Augmented Generation (RAG)* pipeline[3]:

1) **Chunking.**
Abstracts are split into chunks of ≤ 200 tokens with an overlap of 64 tokens, preventing loss of cross-sentence semantics.

2) **Embedding & Indexing.**
Each chunk is encoded with `BAAI/bge-large-en-v1.5` (BGE). The dense vectors, along with raw text and metadata, are stored in a local Chroma vector database.
3) **Initial Retrieval.**
A query is embedded by the same BGE model; cosine similarity returns the Top–k chunks. We tuned $k \in \{13, 10, 8, 6\}$ to balance recall and latency.
4) **Re-ranking.**
The retrieved set is re-scored with a cross-encoder reranker (`BAAI/bge-reranker-base`). The Top–4 chunks by reranker score are selected.
5) **Answer Generation.**
We employ `Mistral-7B-Instruct` as the primary generator and `LLaMA-2-7B-chat` for ablation. A prompt template enforces: "*Answer strictly from provided context; no external information;* $\leq$ 100 *words*," ensuring high-precision medical answers.

III. Metric Definitions and Interpretation

[4] We evaluate the system along two complementary axes: *retrieval quality* and *generation quality.* Retrieval is assessed with **Context Precision** and **Context Recall**; generation is assessed with **Answer Relevancy** and **Faithfulness**. Formal definitions and intuitive explanations follow.

A. Retrieval Metrics

a) Context Precision@K.:

$$\text{CPrec}\,@K = \frac{\sum_{k=1}^{K}\left(\text{Precision}\,@k \times v_k\right)}{\#\text{ relevant items in the top } K \text{ results}} \tag{1}$$

where

$$\text{Precision}\,@k = \frac{\text{true positives@}k}{\text{true positives@}k + \text{false positives@}k}. \tag{2}$$

Interpretation. Context Precision measures how much of the retrieved context is actually useful. It ignores missing evidence and penalises irrelevant or redundant passages; the more focused the retrieval, the higher the score.

b) Context Recall.:

$$\text{CRecall} = \frac{\#\text{ Supported Claims}}{\#\text{ Total Claims}} \tag{3}$$

Interpretation. Context Recall asks whether the retrieval covers *all* evidence needed to answer. If every factual element in the reference answer can be traced back to retrieved passages, recall reaches 1.0.

B. Generation Metrics

a) Answer Relevancy.: Following ragas, we compute the semantic similarity between the system answer and the reference answer with a sentence–embedding model. A score of 1.0 indicates perfect semantic overlap.

b) Faithfulness.:

$$\text{Faithfulness} = \frac{\#\,\text{Answer claims supported by retrieved}}{\#\,\text{Total claims in the answer}} \tag{4}$$

Interpretation. Faithfulness measures factual consistency between the generated answer and the retrieved evidence; it detects hallucinations introduced by the generator.

C. Experimental Observation

In our four–question experiment, both *Context Precision* and *Context Recall* reach 1.0 (Table I), indicating that the agent retrieved exactly the information required—nothing more, nothing less. This suggests that even with a 7 B-parameter backbone, the agent fully utilises the 2 400-document corpus. Generation metrics (Table II) show high relevancy overall but reveal faithfulness drops on two questions, pointing to numeric hallucination at the generation stage.

IV. Results and Interpretation

A. Retrieval Results

TABLE I
Retrieval metrics on 4-question sample

Question	**Precision**	**Recall**
Q1	1.00	1.00
Q2	1.00	1.00
Q3	1.00	1.00
Q4	1.00	1.00

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

TABLE II
Generation metrics on 4-question sample

Question	**Faithfulness**	**Answer Relevancy**
Q1	0.89	1.00
Q2	0.88	0.93
Q3	0.57	0.00
Q4	0.54	0.96

a) Qualitative Analysis.: To understand how the numerical metrics translate into real behaviour, we analysed four representative queries covering mechanism comparison, negation, neuro-imaging evidence synthesis, and cross-domain catalysis:

Q1 *"How does sonodynamic therapy (SDT) differ from conventional antibiotics in combating multidrug-resistant infections?"*
Faithfulness 0.89 — Relevancy 1.00. The answer correctly contrasted ROS-mediated damage with antibiotic resistance pathways, fully supported by context. *Insight:* the agent excels at mechanism-based biomedical explanations.

Q2 *"Does zebrafish possess an ortholog/paralog of mammalian calprotectin?"*
Faithfulness 0.88 — Relevancy 0.93. The model produced the correct negative conclusion while preserving mechanistic detail, showing it can handle negation and complex homology reasoning.

Q3 *"What is the current understanding of gray-matter alterations in vestibular migraine?"*
Faithfulness 0.57 — Relevancy 0.00. The response wandered into unrelated depression literature; only partial claims were context-supported. *Diagnosis:* recall was perfect, yet reranking failed to surface the most pertinent VBM passages, and the generator hallucinated unsupported claims.

Q4 *"How do nickel clusters improve the hydrogen-evolution reaction (HER)?"*
Faithfulness 0.54 — Relevancy 0.96. Core catalytic mechanisms were captured, but half the mechanistic claims lacked explicit support. *Lesson:* cross-disciplinary prompts are answered fluently, but require stricter citation control.

b) Take-aways.:

- Perfect retrieval (**precision = recall = 1.0**) guarantees the evidence pool is complete, yet generation may still hallucinate — explaining the divergence between Context Recall and Faithfulness.
- High Relevancy (≥ 0.93 on three queries) confirms intent alignment, but Faithfulness exposes factual fragility when the generator reformulates numeric or domain-specific detail.
- Failure analysis (Q3) suggests reranker weight tuning and citation-aware decoding as immediate avenues for improvement.

c) Implications.: Perfect retrieval ensures clinicians are not burdened by irrelevant literature, but generation quality still limits direct clinical adoption. We plan to add citation-aware decoding and numeric consistency checks to mitigate remaining hallucinations (see Future Work).

V. Conclusion

We presented Agent-MedRAG, a domain-specific retrieval-augmented agent that achieves **perfect** context precision and recall on a 2 400-document PubMed subset using only a 7-B-parameter backbone. Generation quality is generally high (answer relevancy ≥ 0.93 on three of four queries), but faithfulness analysis reveals residual hallucination when numeric or domain-specific details are reformulated. Overall, the system demonstrates that lightweight RAG pipelines can deliver evidence-grounded biomedical answers, with hallucination control emerging as the primary path for improvement.

VI. Future Work

1) **Scaling to Larger Language Models for Improved Clinical Reasoning**
Integrate stronger backbones such as LLaMA-3-70B or GPT-4–class models to enhance mechanistic reasoning and interpretability in complex contexts.
2) **Conversational Memory and Multi-Turn Questioning**
Add a memory module and dialogue interface so users can iteratively refine queries, progressively converging on precise clinical answers.
3) **Incremental Indexing and Robustness to New Literature**
Implement streaming ingestion and dynamic re-indexing so newly published studies are assimilated without full re-embedding, keeping knowledge fresh.
4) **Evidence Grading and Literature Prioritisation**
Incorporate an evidence-grading framework (e.g. GRADE) to rank retrieved literature by confidence level, giving clinicians a clear quality signal.

References

[1] National Library of Medicine. (n.d.) Pubmed. U.S. National Library of Medicine. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/

[2] Mistral AI. (2023, Sep.) Announcing mistral 7b. [Online]. Available: https://mistral.ai/news/announcing-mistral-7b

[3] ""rag ragas "," https://zhuanlan.zhihu.com/p/676192377, 2025, accessed 15 May 2025.

[4] Ragas. (2025) Available metrics. [Online]. Available: https://docs.ragas.io/en/latest/concepts/metrics/available_metrics/