Comparative Evaluation of RAG vs Base LLM

1. Objective

The objective of this evaluation was to determine whether a Retrieval-Augmented Generation (RAG) system produces more reliable and clinically relevant responses than a standalone Large Language Model (LLM).

The hypothesis was that incorporating retrieval from historical medical cases would improve answer relevance, contextual grounding, and overall reliability compared to a base LLM operating without retrieval.

2. Experimental Setup

Both systems were evaluated under identical generation settings:

- Model: Mistral (local via Ollama)
- Temperature: 0.0 (deterministic output)
- Same token limits and response style
- Only difference: RAG included retrieved case context

The RAG pipeline used:

- Recursive chunking with overlap
- Hybrid retrieval (Dense + BM25)
- FAISS vector database
- MPNet embeddings
- Structured prompt engineering

Five clinically oriented queries were used for comparison.

3. Evaluation Metrics

Four complementary metrics were used:

3.1 Answer Relevance (LLM as a judge)

Measures how directly the response addresses the clinical query.
Scale: 1 (low) to 5 (high).

3.2 Grounding / Faithfulness (LLM as a judge)

Measures whether the response is supported by retrieved cases.
Scale: 0 (unsupported) to 2 (fully grounded).

3.3 Hallucination Rate (LLM as a judge)

Binary evaluation of whether unsupported diagnoses or findings were introduced.

3.4 Semantic Similarity (Automatic – BERTScore)

Measures conceptual similarity between generated answers and retrieved context.
This metric captures semantic overlap but does not guarantee factual correctness.

4. Results Summary

Metrics	Base LLM	RAG
Avg Relevance	2.2	4.2
Avg Grounding	1.8	2.8
Hallucination Rate	20%	0%
Avg Semantic Similarity	~0.80	~0.83

5. Analysis

The base LLM frequently provided conservative responses such as "Insufficient information." While this minimized hallucination, it reduced clinical usefulness and relevance.

In contrast, the RAG system generated structured, case-based summaries grounded in historical medical records. It consistently demonstrated higher relevance and full contextual grounding.

Semantic similarity scores were sometimes close between the two systems. This highlights a limitation of automatic metrics: semantic alignment alone does not distinguish between generic medical language and evidence-backed reasoning. Manual evaluation was essential to capture this difference.

RAG responses required additional computation time due to retrieval, representing a trade-off between speed and reliability.

6. Conclusion

The evaluation demonstrates that Retrieval-Augmented Generation improves clinical relevance and contextual grounding compared to a standalone LLM.

While the base model avoided hallucination by remaining cautious, it lacked specificity. The RAG system provided more structured, evidence-aligned responses, making it better suited for domain-sensitive applications such as healthcare.

Detailed responses, individual scores, and per-query evaluations are available in the `evaluation_results.csv` file.