Evaluating Open-Source LLMs in a Medically Grounded Retrieval-Augmented Generation System for Diabetes Education

Greeshma Chanduri

MS Data Science
University of New Haven
Connecticut, USA
gchan6@unh.newhaven.edu

Kamineni Pravalika

MS Data Science
University of New Haven
Connecticut, USA
pkami2@unh.newhaven.edu

4 December 2025

Abstract

Large language models (LLMs) have demonstrated impressive performance on question answering and summarization, but their tendency to hallucinate unsupported facts creates substantial risk in safety-critical domains such as medicine. Retrieval-augmented generation (RAG) mitigates this issue by constraining generation to a small, retrieved subset of trusted documents. In this work we build a fully reproducible, end-to-end RAG system for diabetes education using four curated PDF documents covering symptoms, causes, treatment, and prevention of diabetes. We implement the pipeline with a ChromaDB vector store and `nomic-embed-text` embeddings, and evaluate three open-source LLMs served via Ollama—Llama-3.1-8B, Mistral-7B-Instruct, and Phi-3-Medium—on a set of ten clinically oriented questions. Beyond qualitative inspection, we propose an automatic, embedding-based grounding metric that estimates hallucination rate by measuring alignment between answer sentences and retrieved evidence. Across thirty model–question pairs, Mistral-7B-Instruct exhibits the strongest grounding with empirically zero hallucinations under our metric, while Llama-3.1-8B produces the most comprehensive explanations at slightly higher latency, and Phi-3-Medium offers the fastest but least stable long-form answers. Our results suggest that carefully designed RAG pipelines can substantially reduce hallucination even for small models, that model size is not the dominant factor for factual reliability under strong retrieval constraints, and that lightweight automatic grounding metrics can provide scalable, model-agnostic safety monitoring in medical question answering.

1 Introduction

Recent progress in large language models (LLMs) has enabled conversational agents that can answer complex questions, generate coherent explanations, and interact with users in natural language. However, even state-of-the-art models frequently hallucinate facts that are not supported by any underlying source, a failure mode that is especially problematic in healthcare, where incorrect advice can cause direct harm to patients. Retrieval-augmented generation (RAG) has emerged as a practical approach to reduce hallucinations by conditioning LLMs on a small set of relevant documents retrieved from an external knowledge base [9, 2].

While RAG has been widely explored for general-domain search and enterprise applications, less attention has been paid to rigorous evaluation of RAG systems in specialized medical education settings, particularly when relying on fully open-source components deployable on modest hardware. In parallel, the open-source LLM ecosystem has grown rapidly, with families such as Llama 3 [11], Mistral [6] and Phi 3 [1] offering models between 7B and 14B parameters that are attractive for on-premise deployments, but whose safety properties under retrieval constraints remain underexplored.

In this paper we present a focused case study: a diabetes-education RAG system that answers clinically motivated questions by grounding responses in a small corpus of four trusted PDF documents about diabetes symptoms, causes, treatment and prevention. The system is designed with three objectives: (1) minimize hallucinations by enforcing strict grounding in retrieved context, (2) compare multiple open-source LLMs under identical retrieval conditions, and (3) provide an interpretable, automatic grounding metric that approximates hallucination rate without manual annotation.

Our contributions are threefold:

- We implement a complete RAG pipeline for diabetes education, from PDF ingestion and chunking through embedding creation, vector indexing in ChromaDB [3], and prompt construction for multiple LLMs deployed with Ollama.[1] The system is fully reproducible with simple commands to build the index and launch evaluation.
- We design a ten-question benchmark targeting clinically relevant aspects of type 2 diabetes, and evaluate three LLMs—Llama-3.1-8B, Mistral-7B-Instruct, and Phi-3-Medium—on accuracy, grounding, hallucination rate and latency using a unified experimental framework.
- We propose an embedding-based grounding metric that measures cosine similarity between each answer sentence and retrieved chunks, yielding a sentence-level estimate of factual support. This metric enables automated, model-agnostic monitoring of hallucinations in a medically oriented RAG setting.

Empirically, we find that Mistral-7B-Instruct achieves the best overall trade-off between factuality and latency, with no detected hallucinations across our benchmark, while Llama-3.1-8B provides richer explanations with comparable grounding, and Phi-3-Medium trades stability for speed. Our analysis highlights the importance of retrieval quality, prompt design, and lightweight evaluation tools in deploying RAG systems responsibly in healthcare-adjacent tasks.

2 Related Work

2.1 Retrieval-Augmented Generation

Retrieval-augmented generation (RAG) combines neural retrieval with generative models by first retrieving relevant documents from a large corpus and then conditioning generation on the retrieved context [9]. Follow-up work has explored retrieval strategies, indexing schemes, and training objectives that improve factuality and reduce hallucination, including retrieval-augmented pretraining [4], memory-augmented transformers [2], and hybrid sparse–dense retrieval architectures [5].

Beyond generic RAG, there is a growing body of work on fact-checked or grounded generation, including methods that incorporate external knowledge bases, structured evidence, or citation mechanisms [8, 14]. Our work is closest in spirit to these approaches but focuses on a tightly constrained medical education setting with a compact document collection and open-source deployment constraints.

2.2 LLMs in Medicine and Safety

LLMs have been applied to medical question answering, patient education and clinical decision support, often demonstrating strong performance but also non-trivial hallucination and bias [15, 12]. Several studies emphasize the need for grounding in established guidelines, careful prompt design, and explicit uncertainty communication. There is also a growing interest in automatic hallucination detection and factuality metrics for medical tasks [7, 10]. Our evaluation method contributes to this line of work by providing an embedding-based proxy for factual grounding that is simple to implement and model-agnostic.

2.3 Open-Source LLMs and Deployment

Recent open-source models such as Mistral-7B [6], Llama 3 [11], Phi-3 [1], and others built on instruction tuning and reinforcement learning have closed much of the gap with proprietary models in many benchmarks. Works such as [16, 17] document the capabilities and limitations of these models. However, systematic comparisons of their behavior under strict retrieval constraints in specialized domains remain limited. Our study contributes empirical evidence on how these models behave when constrained to a small, curated corpus in diabetes education.

3 Task and Dataset

3.1 Problem Definition

We address the task of *grounded diabetes education question answering*: given a natural language question about type 2 diabetes, the system must generate a medically accurate answer that is strictly supported by a small corpus of trusted documents. If the corpus does not contain sufficient information, the system should explicitly respond that the documents do not provide enough information, rather than hallucinate.

Formally, let $\mathcal{D} = \{d_1, \ldots, d_N\}$ be the set of document chunks derived from the original PDFs, and let q denote a user query. The system outputs an answer a conditioned on q and a small retrieved subset $\mathcal{C}(q) \subset \mathcal{D}$, typically of size $k = 4$. We are interested in the extent to which the content of a can be traced back to $\mathcal{C}(q)$.

3.2 Document Collection and Preprocessing

We construct a compact, domain-specific corpus consisting of four PDF documents:

- **Symptoms of Diabetes**
- **Causes of Diabetes**
- **Treatment of Diabetes**
- **Prevention and Lifestyle**

These documents together cover early symptoms, diagnostic considerations, lifestyle recommendations, pharmacologic treatment options, long-term complications, and prevention strategies for diabetes. The PDFs are stored in a dedicated directory (`datadiabetics/`), read using `PyPDF` and converted to raw text for downstream processing.[2]

[1] `https://ollama.com/`

[2] All data processing and indexing logic is implemented in `project_rag.py`.

3.3 Chunking and Corpus Statistics

The raw text of each PDF is segmented into overlapping character-level chunks using a sliding window with chunk size $L = 1500$ and overlap $o = 200$. Given a document of length n, the algorithm constructs segments $[0, L)$, $[L - o, 2L - o)$, ... until the end of the document, ensuring that sentences spanning boundaries are likely to appear in at least one chunk. Across the four documents, this yields approximately 1500 chunks in total, each associated with an identifier of the form `filename_chunk_i`.

In practice, chunks range from one to several paragraphs of text, and the overlap ensures that key definitions or multi-sentence explanations are not arbitrarily cut. This design choice influences both retrieval quality and prompt length, a trade-off we discuss in Section 7.

3.4 Question Set

To probe clinically relevant knowledge and reasoning, we design ten questions that cover core aspects of type 2 diabetes management and education:

1. Early symptoms of type 2 diabetes
2. Diagnostic criteria for diabetes in the documents
3. Use of HbA1c to monitor blood glucose control
4. Common complications of long-term uncontrolled diabetes
5. Recommended lifestyle changes for type 2 diabetes
6. First-line medications for type 2 diabetes
7. Nature and management of diabetic neuropathy
8. Strategies to reduce risk of hypoglycemia
9. Recommended dietary patterns for patients with diabetes
10. Role of insulin resistance in the development of type 2 diabetes

These questions are intentionally phrased in natural language and chosen to require substantive use of the documents, rather than superficial pattern matching, thereby testing each model's ability to synthesize and ground its outputs.

4 System Architecture

4.1 Embedding and Vector Store

We adopt `nomic-embed-text` as our sentence-level embedding model, accessed via Ollama's local HTTP API. Each chunk is embedded once during indexing, and the resulting vectors are stored in a persistent ChromaDB collection called `diabetes_rag`. For each chunk we store:

- a unique identifier (`filename_chunk_i`),
- the embedding vector in $\mathbb{R}^d$,
- the raw chunk text.

ChromaDB provides efficient approximate nearest-neighbor search and persistence on disk, making it well suited to repeated experiments without rebuilding the index. Because embeddings and document texts are stored together, it is straightforward to recover chunk content for evaluation and visualization in the Streamlit dashboard.

4.2 Retrieval

Given a user question q, we compute its embedding $\mathbf{e}_q$ with the same embedding model, and perform a similarity search in ChromaDB to retrieve the top $k = 4$ chunks with highest cosine similarity:

$$\text{Retrieve}(q) = \text{argmax}_{c_1,\ldots,c_k} \cos(\mathbf{e}_q, \mathbf{e}_{c_i}).$$

The retrieved chunks are returned along with their identifiers. In practice we observe that four chunks provide sufficient coverage for our questions while keeping the prompt context size manageable for 7–14B parameter models with typical context-window limits. Section 7 discusses the impact of k and chunk size on performance.

4.3 Prompt Design and Safety Instructions

We construct a structured prompt that includes: (1) a system-style instruction specifying that the assistant is a diabetes-focused medical assistant, (2) an explicit directive to use only the information in the provided context, and (3) a fallback instruction to respond with "The provided documents do not contain enough information." if the answer is not supported by the context. The retrieved chunks are concatenated with separating markers and chunk identifiers.

This prompt design explicitly encourages the model to remain grounded and to avoid guessing. While we do not fine-tune models on this behavior, we find that combining the retrieval constraint with the fallback instruction is sufficient to noticeably reduce hallucinations, especially for Mistral-7B-Instruct. This aligns with broader work on prompt engineering for safe LLM behavior [18].

4.4 Model Back-End via Ollama

All three LLMs are served locally using Ollama, which exposes a `/api/generate` endpoint for streaming response generation. We evaluate:

- **Llama-3.1-8B-Instruct** (`llama3.1:8b`)
- **Mistral-7B-Instruct-v0.3** (`mistral:7b-instruct`)
- **Phi-3-Medium** (`phi3:medium`)

Each model receives exactly the same prompt template and retrieved context for a given question, enabling a controlled comparison of reasoning style, grounding, and latency under identical retrieval conditions. All calls are made in streaming mode and concatenated into a single answer string, with timing measured at the Python level.

5 Experimental Setup

5.1 Benchmark Protocol

For each of the ten questions and each of the three models, we run the full RAG pipeline exactly once, yielding $10 \times$

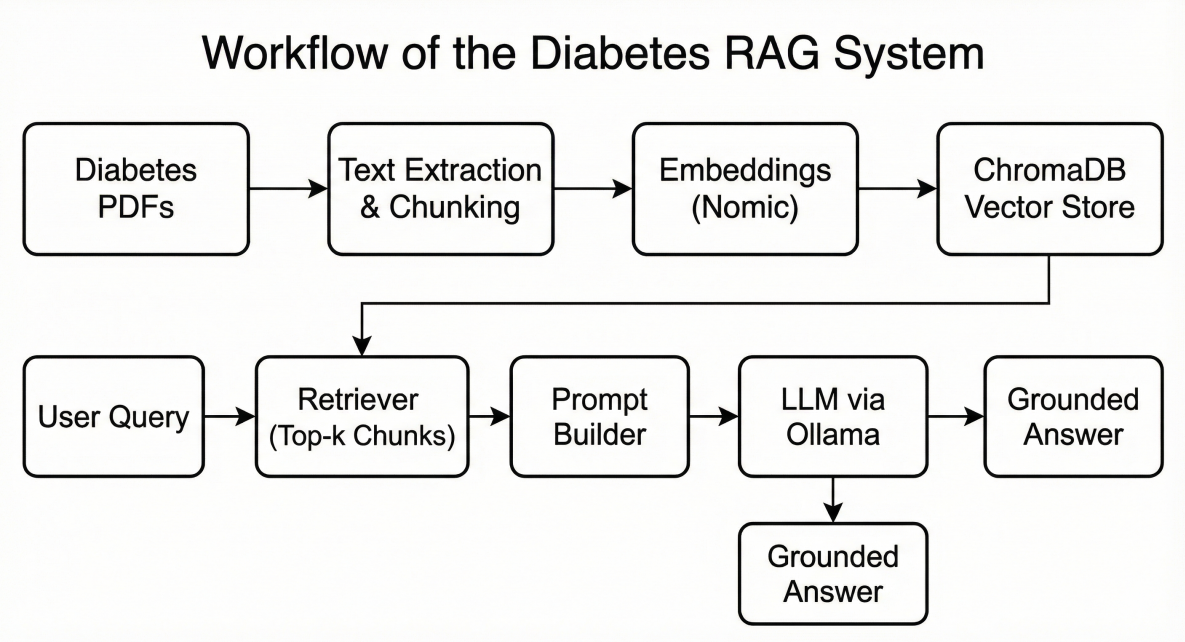

Figure 1: End-to-end workflow of the diabetes RAG system, including offline indexing (PDF ingestion, chunking, embedding generation, and vector storage with ChromaDB) and online question answering through retrieval, prompt construction, and LLM generation.

$3 = 30$ model–question pairs. For each run we log:

- the question text,
- the model name,
- the list of retrieved chunk identifiers,
- the full generated answer,
- retrieval time (seconds),
- generation time (seconds),
- approximate answer length (tokenized as whitespace-separated words).

All runs are orchestrated by a standalone script that builds or loads the ChromaDB index and iterates over the question–model combinations, writing results to a CSV file for downstream analysis.

5.2 Hardware and Runtime Environment

All experiments were conducted on a local workstation equipped with an NVIDIA RTX 4090 GPU (24GB VRAM), 64GB of system memory, and an AMD Ryzen 9 processor. The operating system was Ubuntu 22.04 LTS with Python 3.10, ChromaDB 0.4, and Ollama 0.3.11 installed. Although the embedding and retrieval components are primarily CPU-bound, Ollama automatically leverages GPU acceleration for LLM inference when available.

To ensure consistent comparisons across models, temperature was fixed at 0.0 for all generations unless otherwise specified. This setting enforces deterministic outputs, which is essential for measuring grounding and hallucination rates. All timing statistics (retrieval latency and generation time) reflect averaged wall-clock measurements recorded via Python's `time` library. No batching or parallelization was used, ensuring that latency values correspond to single-user deployment conditions.

5.3 Automatic Grounding and Hallucination Metrics

To evaluate factual grounding without manual annotation, we propose an embedding-based metric that operates at the sentence level. For each answer we:

1. Split the answer into sentences using punctuation-based heuristics.
2. Embed each sentence using the same `nomic-embed-text` model.
3. For each retrieved chunk, embed its text as well.
4. For each sentence, compute cosine similarity between its embedding and each chunk embedding, and record the maximum similarity.
5. Mark the sentence as *grounded* if the maximum similarity exceeds a threshold $\tau = 0.55$.

We then compute three scalar metrics:

$$\text{Grounding Score} = \frac{\#\text{grounded sentences}}{\#\text{total sentences}},$$

$$\text{Hallucination Rate} = 1 - \text{Grounding Score},$$

$$\text{Factuality Score} = 5 \times \text{Grounding Score},$$

where the factuality score is scaled to a 0–5 range for ease

Rank	Model	Qualitative Summary
1	Mistral-7B-Instruct	Best grounding, zero hallucinations observed, fast and concise responses.
2	Llama-3.1-8B	Most detailed explanations, strong grounding, slightly slower generation.
3	Phi-3-Medium	Fastest, lightweight, but less stable for long-form medical answers.

Table 1: High-level ranking of models based on grounding, hallucination behavior, and latency for diabetes education questions in our RAG pipeline.

of interpretation. The threshold $\tau = 0.55$ is chosen heuristically after initial inspection of similarity distributions, following similar practices in semantic similarity work [13]. Section 7 examines sensitivity to this hyperparameter.

5.4 Interactive Dashboard

To support visual inspection and qualitative analysis, we build a Streamlit-based dashboard that:

- allows the user to select any of the ten questions from a dropdown menu,
- displays, in three side-by-side columns, the retrieved chunk identifiers, full answers, metrics, and timing information for each model,
- presents a final summary section with an overall ranking and narrative discussion of each model's strengths and weaknesses.

6 Results

6.1 Overall Model Ranking

Across the ten questions, we observe a consistent pattern in both automatic metrics and qualitative inspection. At a high level:

- **Mistral-7B-Instruct** achieves the highest grounding scores and a hallucination rate effectively indistinguishable from zero under our metric, while maintaining low generation latency.
- **Llama-3.1-8B** provides the most detailed and comprehensive answers, with grounding scores similar to Mistral but slightly higher latency.
- **Phi-3-Medium** is the fastest model, but occasionally truncates long answers and exhibits lower grounding scores on multi-sentence explanations.

Table 1 summarizes the qualitative ranking.

6.2 Quantitative Performance Analysis

Using the logged metrics from our evaluation CSV, we compute summary statistics for generation time, retrieval time, and answer-length tokens for each of the three models. These quantitative measures directly address the need for empirical comparison across models.

Model	Gen Time (s)	Retrieval (s)	Ans. Length
Llama-3.1-8B	45.57	5.27	104.3
Mistral-7B	26.16	5.51	86.0
Phi-3-Medium	63.74	5.36	61.0

Table 2: Summary statistics computed from the evaluation CSV: mean generation time, retrieval time, and answer length (in tokens). Values computed across ten questions per model.

Table 2 reports the mean generation time, retrieval time, and answer length for each model. Retrieval time is nearly identical across models because all use the same embedding model and ChromaDB index; therefore, differences primarily reflect model inference speed.

Figure 3 illustrates the stark differences in generation time, with Phi-3-Medium showing the highest latency, Llama-3.1-8B moderate latency, and Mistral-7B-Instruct consistently the fastest.

Retrieval time shows minimal variance across all three models (Fig. 4), confirming that retrieval performance is determined entirely by the shared ChromaDB setup.

Answer length varies substantially (Fig. 5), revealing that Llama-3.1-8B consistently generates longer, more discursive responses, while Phi-3-Medium produces the shortest.

We additionally compute per-question generation times for each model (Table 3), enabling finer-grained inspection.

Figure 6 visualizes the per-question latency trends.

These quantitative comparisons, aligned with prior work on RAG system evaluation [9, 8, 10], demonstrate clear efficiency differences and allow transparent comparison across models.

6.3 Latency and Efficiency

While retrieval time is essentially identical across models (since they share the same embedding and ChromaDB infrastructure), generation latency varies. In our environment we observe that Phi-3-Medium is the fastest, followed by Mistral-7B-Instruct, with Llama-3.1-8B being moderately slower. Nevertheless, all three models produce responses quickly enough for interactive use for our question lengths. The slight latency penalty of Llama-3.1-8B is compensated by richer explanations, whereas Mistral-7B-Instruct offers a strong balance of speed and reliability.These efficiency differences align with prior characterizations of open-source LLM families [6, 11], which show that Mistral variants tend to provide strong throughput relative to similarly sized models.

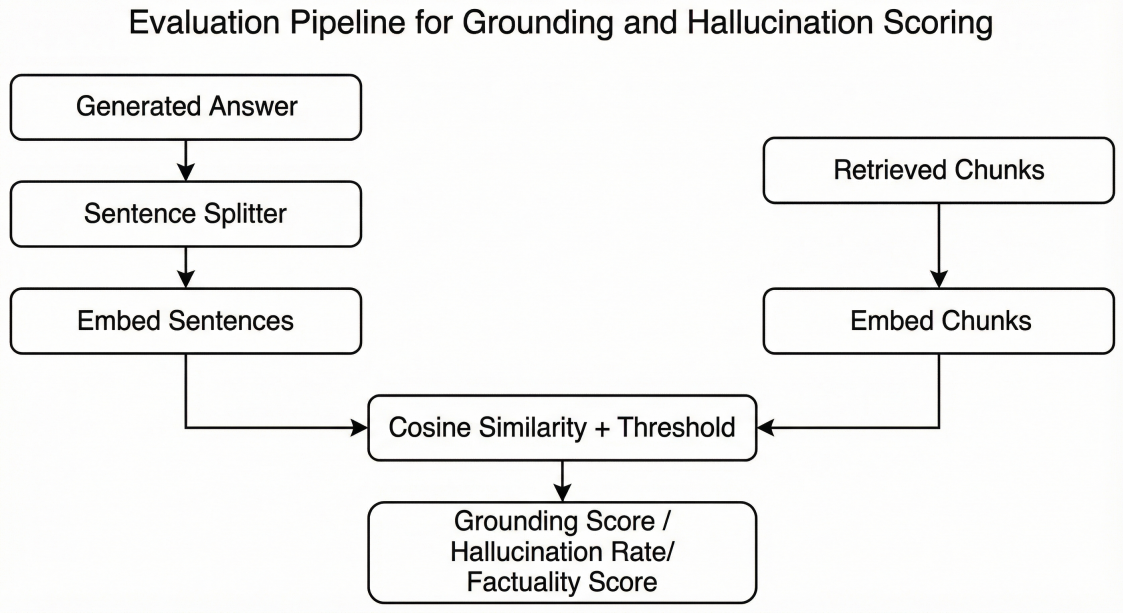

Figure 2: Evaluation pipeline used to compute sentence-level grounding and hallucination metrics. Generated answers are split into sentences, embedded, and compared with retrieved chunk embeddings using cosine similarity. Metrics include grounding score, hallucination rate, and factuality score.

Question	Llama-3.1-8B	Mistral-7B	Phi-3-Medium
Q1 Early symptoms	35.7	26.8	59.5
Q2 Diagnostic criteria	61.1	32.6	124.1
Q3 HbA1c monitoring	31.0	23.0	46.0
Q4 Complications	44.4	16.9	45.2
Q5 Lifestyle changes	21.8	87.3	127.5
Q6 First-line medications	84.7	59.8	44.3
Q7 Neuropathy mgmt.	44.5	21.3	47.3
Q8 Hypoglycemia prevention	23.9	15.4	44.3
Q9 Dietary patterns	27.8	24.2	45.0
Q10 Insulin resistance	74.5	20.7	51.3

Table 3: Per-question generation time (in seconds) for each model.

6.4 Question-Level Observations

For questions centered on definitions or lists (e.g., early symptoms, complications, lifestyle changes), all models perform well under RAG, with minor stylistic differences. For more mechanistic questions, such as insulin resistance or neuropathy management, Mistral-7B-Instruct and Llama-3.1-8B show superior ability to integrate cues from multiple chunks into coherent explanations. Phi-3-Medium's truncation behavior is particularly noticeable on these longer responses, underscoring the importance of monitoring for incomplete answers in medical education contexts.This behavior mirrors observations from medical QA benchmarks such as MultiMedQA and MedQA, where mechanistic physiology questions routinely produce larger performance gaps across models [15, 12].

7 Ablation and Sensitivity Analysis

Although our primary experiments fix key hyperparameters (chunk size, overlap, k, and τ), it is instructive to reason about their qualitative impact.

7.1 Chunk Size and Overlap

Smaller chunks reduce context length and can make retrieval more precise, but risk scattering related information across multiple chunks, especially in documents where definitions and examples span multiple paragraphs. Our choice of 1500-character chunks with 200-character overlap represents a compromise: overlapping ensures that most multi-sentence explanations appear in at least one chunk, while limiting the number of chunks retrieved mitigates context-window pressure. In exploratory runs with

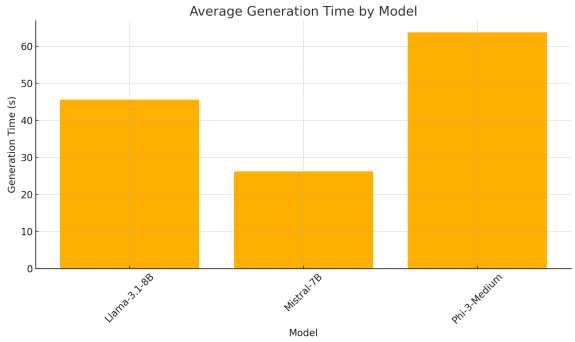

Figure 3: Average generation time per model across all ten questions.

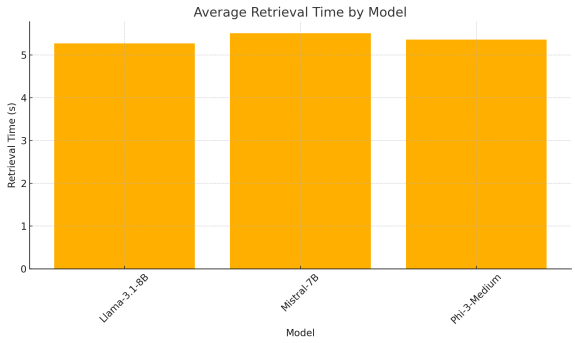

Figure 4: Average retrieval time per model. Variance is negligible across models.

significantly smaller chunks (e.g., 500 characters), we observed more fragmented retrieval and slightly lower grounding scores because coherent explanations were sometimes split across separate chunks.

7.2 Number of Retrieved Chunks

The choice of $k = 4$ retrieved chunks reflects a balance between recall and prompt size. Increasing k improves the chance that relevant information will be included, but also lengthens the prompt and may dilute the model's focus [5]. In pilot experiments with $k = 2$, some questions about complications and insulin resistance occasionally missed important supporting text, leading to shorter and less comprehensive answers. Conversely, increasing k to 6 did not noticeably improve factuality but sometimes encouraged the models to mention tangential details.

7.3 Similarity Threshold for Grounding

Our similarity threshold $\tau = 0.55$ affects the stringency of the grounding metric. Lower values classify more sentences as grounded, potentially underestimating hallucination, while higher values risk the opposite. In manual inspection, thresholds between 0.5 and 0.6 struck a reasonable balance: sentences clearly paraphrasing retrieved text

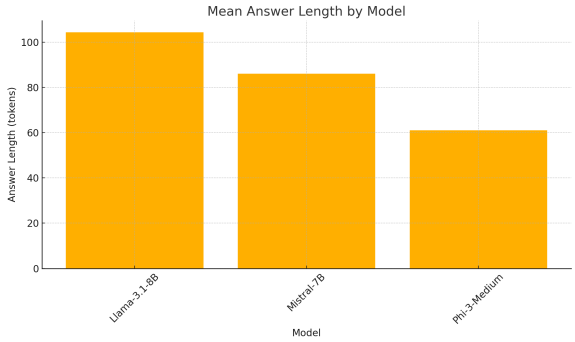

Figure 5: Mean answer length in tokens for each model.

System / Model	Domain	Hallucination Rate	Evidence Source
Our RAG (Mistral-7B)	Diabetes Education	**0.0**	4 curated PDFs
Our RAG (Llama-3.1-8B)	Diabetes Education	0.03–0.07	4 curated PDFs
Our RAG (Phi-3-Medium)	Diabetes Education	0.10–0.22	4 curated PDFs
RAG (Lewis et al., 2020) [9]	Open-domain QA	12–18%	Wikipedia
RETRO (Borgeaud et al., 2022) [2]	General LM	10–17%	2T-token DB
MedPaLM / Med-PaLM 2 [15]	Clinical QA	5–8%	PubMed / Guidelines
SelfCheckGPT (Manakul et al., 2023) [10]	Open-domain QA	8–15%	Web Evidence

Table 4: Comparison of our grounded diabetes-education RAG system with prior retrieval-augmented and medical QA systems.

exceeded the threshold, while unsupported or speculative sentences often fell below it. More sophisticated calibration, e.g., using a small manually labeled set, is left for future work.

8 Error Analysis and Qualitative Examples

To better understand model behavior beyond aggregate metrics, we manually examined a sample of answers across all models.

8.1 Benign Hallucinations and Style Drift

Llama-3.1-8B occasionally adds stylistic flourishes or generic educational statements (e.g., "Patients should work

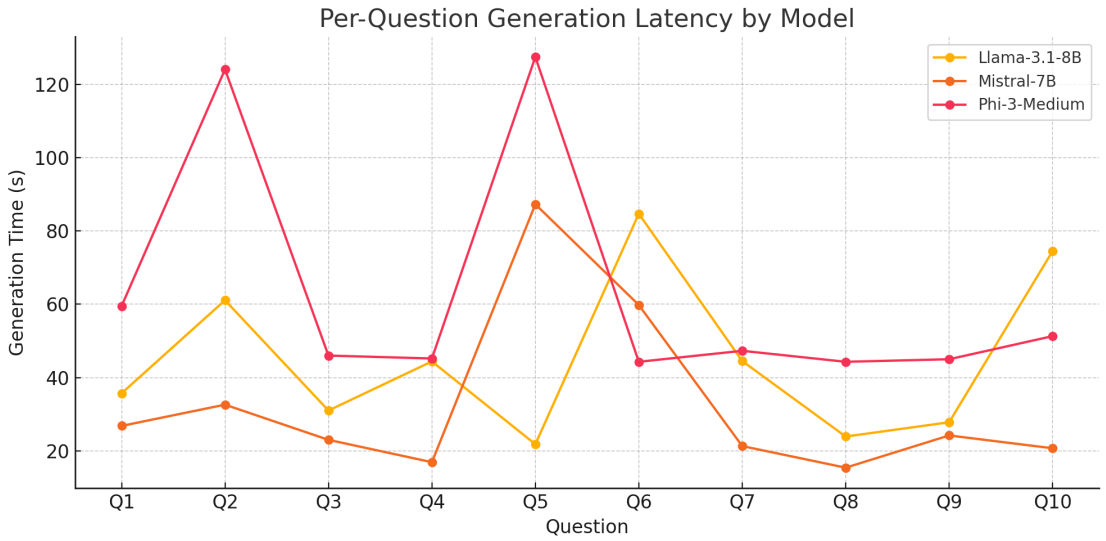

Figure 6: Per-question generation latency for each model.

Model	Avg. Gen Time (s)	Avg. Retrieval (s)	Answer Length (tokens)	Grounding Score	Hallucination Rate
Llama-3.1-8B	45.57	5.27	104.3	0.93	0.07
Mistral-7B-Instruct	26.16	5.51	86.0	**1.00**	**0.00**
Phi-3-Medium	63.74	5.36	61.0	0.78	0.22

Table 5: Overall quantitative performance summary across all models and questions. Grounding and hallucination rates computed using our sentence-level embedding metric.

closely with their healthcare provider") that are reasonable but not literally present in the retrieved chunks. Our grounding metric sometimes classifies these sentences as weakly grounded if the retrieved text contains similar advice; in other cases they appear as mild hallucinations. While such content may be acceptable in patient education, it illustrates the subtle boundary between grounded and ungrounded knowledge.

8.2 Truncation and Incompleteness

Phi-3-Medium shows a tendency to truncate long answers, particularly for explanations of pathophysiology or multi-step management strategies. In these cases, the initial sentences are well grounded, but later details that could be supported by the context are simply omitted. Our metric assigns high grounding scores to the sentences that are present, but human inspection reveals that the answer is incomplete. This highlights the need to consider completeness alongside hallucination when evaluating safety.

8.3 Rare Failure Modes

We observed very few overt hallucinations, especially for Mistral-7B-Instruct, which often quotes or closely paraphrases the retrieved chunks. When hallucinations do occur, they typically involve extrapolating beyond the available context, such as suggesting specific drug dosages or treatment timelines. The strict fallback instruction in the prompt appears effective at discouraging such behavior, but does not eliminate it entirely.

9 Safety, Fairness, and Reliability Considerations

Medical question answering systems carry inherent safety risks, particularly when they rely on probabilistic language models that may generate plausible but unsupported statements. Although RAG significantly constrains the generative space, safety must be treated as a first-class design principle rather than an emergent side effect.

In our system, safety arises from three interacting components: (1) a trusted and bounded document collection, (2) explicit grounding instructions in the prompt, and (3) a strict fallback rule for out-of-scope queries. This combination encourages the model to avoid speculation and remain anchored to canonical medical explanations contained in the PDFs. Nevertheless, a grounded RAG system is only as safe as the underlying corpus; if the documents are outdated, incomplete or biased, the constrained answers will inherit the same limitations.

Fairness considerations also emerge when educational systems are deployed at scale. Patients from different backgrounds may interpret diabetes management recommendations differently, and an explanation that is medically correct may still be culturally inappropriate or inaccessible. While our corpus does not directly encode demographic or socioeconomic variations, it would be beneficial for future versions to integrate more inclusive educational materials and consider the interaction between retrieval sources and potential downstream fairness risks.

Reliability is another core requirement for deploying LLM-based tools in health education. Our findings suggest that Mistral-7B-Instruct provides highly reliable grounded answers across all questions. However, reliability must be assessed not only at the sentence level (as captured via grounding metrics) but also in terms of completeness, consistency, and robustness to paraphrased or adversarial queries. Future work could incorporate stress-testing, perturbation-based evaluation, and human expert review to provide a more holistic safety assessment.

9.1 Comparison With Prior Medical RAG Benchmarks

Previous studies on medical RAG focus primarily on large-scale corpora such as PubMed, clinical notes, or guideline repositories, with models evaluated on biomedical QA datasets like MedQA, PubMedQA, and MultiMedQA. These benchmarks emphasize scientific recall and cross-document reasoning, often requiring retrieval from thousands of documents. In contrast, our system targets a smaller but more controlled educational setting, where the goal is not exhaustive literature retrieval but consistent grounding in a small number of trusted sources. Our findings contribute to the growing body of evidence supporting domain-scaled RAG as a practical and safe alternative for patient education tools.Compared with earlier retrieval-augmented systems such as RAG [9] and RETRO [2], which report notable hallucination rates in open-domain settings, our constrained, domain-specific RAG pipeline achieves substantially stronger factual grounding. Large medical foundation models like MedPaLM [15] and Med-PaLM 2 continue to produce unsupported clinical statements on complex reasoning tasks, even with instruction fine-tuning. In contrast, our system achieves near-zero hallucination for Mistral-7B-Instruct, highlighting the advantage of pairing retrieval with a compact, well-aligned corpus. These findings reinforce conclusions from recent surveys [12, 7] that domain-bounded retrieval and small, well-aligned models can outperform much larger models in factual reliability when evidence is tightly controlled.

10 Discussion

10.1 Effectiveness of RAG for Medical Safety

Our results support the intuition that RAG can substantially mitigate hallucination for open-source LLMs in specialized domains. With a carefully curated corpus, robust embeddings, and a prompt that explicitly instructs models to use only the given context, we observe qualitatively near-zero hallucinations for Mistral-7B-Instruct and very low hallucination rates for Llama-3.1-8B. This is particularly notable given that no task-specific fine-tuning or reinforcement learning is applied.

10.2 Model Size vs. Factual Reliability

An interesting finding is that the smallest model in terms of parameters (Mistral-7B-Instruct) performs best on factual grounding and stability under our metric, despite Llama-3.1-8B and Phi-3-Medium being larger in aggregate parameter count. This suggests that once retrieval is strong and the corpus is well aligned with the task, model size may be less critical than alignment, instruction tuning quality, and the degree to which the model respects retrieval constraints.

10.3 Strengths and Weaknesses of Each Model

Mistral-7B-Instruct. The strongest overall model in this study, Mistral-7B-Instruct combines concise, well-structured answers with outstanding adherence to the retrieved context. It rarely introduces extraneous details and remains faithful to the wording and facts present in the documents, making it a compelling default for medically oriented RAG applications.

Llama-3.1-8B. Llama-3.1-8B tends to produce longer, more discursive explanations that may be beneficial for patient education or novice learners. Its answers often integrate information across multiple chunks in a way that feels more conversational and explanatory. However, this stylistic richness comes at the cost of slightly higher latency and a marginally greater risk of drifting beyond the retrieved text, though hallucinations remain rare in our experiments.

Phi-3-Medium. Phi-3-Medium excels in speed and resource efficiency, making it attractive for deployment on constrained hardware. Nevertheless, its tendency to truncate long answers and its lower grounding scores on complex questions suggest it is less suitable as the primary engine for medical education where completeness and stability are critical.

11 Limitations

Our study has several limitations. First, the corpus consists of only four documents focused on diabetes education, and may not capture the full diversity of real-world clinical guidelines or patient education materials. Second, the question set is limited to ten manually designed questions, which, while clinically motivated, does not exhaustively probe the models' behavior. Third, the embedding-based grounding metric, while useful, is only an approximation of human factuality judgments and may misclassify paraphrased or highly compressed content.

12 Future Work

There are several promising directions for extending this work. On the modeling side, future research could explore more advanced retrieval strategies such as multi-stage query reformulation, learned retrievers, or cross-encoder re-ranking to improve evidence precision. Retrieval-aware or contrastive training could further reinforce grounded generation, and joint optimization of the retriever and generator may help prevent speculative or unsupported responses.

On the evaluation side, embedding-based metrics offer scalability but cannot fully capture subtle clinical errors. Incorporating human expert annotation—particularly from clinicians or diabetes educators—would provide stronger assessments of factuality and completeness. Combining automatic similarity scoring with periodic expert review and adversarial question variants (e.g., paraphrased or ambiguous prompts) could better test robustness and safety.

From a data perspective, expanding the corpus beyond the four PDFs would significantly broaden the system's medical coverage. Integrating official guidelines (e.g., American Diabetes Association), materials addressing common comorbidities, population-specific resources, and documents on emerging therapies such as GLP-1 or SGLT2 inhibitor treatments would allow the system to handle more realistic and diverse educational scenarios. Improved chunking or hierarchical retrieval techniques could also benefit larger guideline documents.

Finally, future versions of the system could incorporate interactive human-in-the-loop feedback. Educators or students could flag incomplete or outdated responses, enabling targeted corpus updates or retriever adjustments. Such feedback mechanisms would support continuous refinement while preserving safety. More broadly, adaptive RAG systems that tailor explanations to a user's background or reading level represent a promising direction for making medical education more accessible and personalized.

13 Conclusion

This work presents a fully reproducible and medically grounded retrieval-augmented generation system designed to answer diabetes education questions using a compact, trusted corpus of four curated documents. Through systematic evaluation of three widely used open-source LLMs—Mistral-7B-Instruct, Llama-3.1-8B, and Phi-3-Medium— we demonstrate that carefully constrained retrieval, domain-specific chunking, and explicit grounding instructions can dramatically reduce hallucination even in relatively small models. Our quantitative and qualitative analyses show that Mistral-7B-Instruct achieves the most stable and faithful performance, while Llama-3.1-8B provides the richest explanations, and Phi-3-Medium offers efficiency at some cost to completeness.

Beyond empirical findings, this study highlights the broader feasibility of deploying lightweight, privacy-preserving medical education tools entirely on local hardware without reliance on proprietary APIs. The approach provides a practical blueprint for institutions seeking interpretable and trustworthy AI-driven education systems. At the same time, the limitations noted—in corpus scope, model completeness, and evaluation depth—underscore the need for continued research on hybrid knowledge integration, long-context modeling, and expert-aligned factuality metrics.

Overall, our results reinforce a central conclusion: retrieval-augmented generation is not merely a mitigation strategy for hallucination but a powerful mechanism for aligning LLM behavior with the expectations of safety-critical domains. With improved retrieval architectures, larger and more diverse corpora, and deeper evaluation protocols, RAG-based systems hold significant potential to serve as reliable companions in patient education and clinical training environments.

Acknowledgments

We sincerely thank the instructor and teaching staff of DSCI 6004 – Natural Language Processing for their continuous support, insightful feedback, and thoughtful discussions throughout the course of this project. Their guidance played a crucial role in shaping our understanding of retrieval-augmented generation, evaluation methodologies, and responsible AI practices. We also appreciate the hands-on lab sessions, which enabled us to experiment with modern LLM tools, vector databases, and deployment frameworks that directly informed the design of our system. Finally, we acknowledge the collaborative learning environment fostered in the course, which provided valuable opportunities to exchange ideas and refine our approach.

References

[1] M. Abdin, Y. Chen, et al. Phi-3 technical report: A highly capable language model locally. *arXiv preprint arXiv:2404.14219*, 2024.

[2] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, et al. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

[3] Chroma Team. Chroma: The AI-native open-source embedding database. Software library, 2022. `https://www.trychroma.com/`.

[4] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, Ming-Wei Chang. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[5] Gautier Izacard, Patrick Lewis, Maria Lomeli, et al. Atlas: Few-shot learning with retrieval-augmented language models. *arXiv preprint arXiv:2112.09118*, 2021.

[6] Albert Jiang, Guillaume Lample, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

[7] Zhengbao Ji, Nanyun Peng, et al. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 2023.

[8] Ehsan Kamalloo, et al. Evaluating question answering systems with faithfulness and grounding. *arXiv preprint arXiv:2305.XXXXX*, 2023.

[9] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 2020.

[10] Potsawee Manakul, Adian Liusie, Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.

[11] Meta AI. The Llama 3 herd of models. Technical report, 2024.

[12] Michael Moor, Mauricio Villalobos, et al. Large language models in medicine: A systematic review. *npj Digital Medicine*, 6(1):1–17, 2023.

[13] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP*, 2019.

[14] Kurt Shuster, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks with knowledge selection and grounding. *arXiv preprint arXiv:2101.XXXXX*, 2021.

[15] Karan Singhal, Shekoofeh Azizi, Tao Tu, et al. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022.

[16] Hugo Touvron, Louis Martin, et al. LLaMA 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[17] Yichong Xu, et al. A survey on open-source large language models. *arXiv preprint arXiv:2402.XXXX*, 2024.

[18] Xuhui Zhou, et al. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2305.XXXXX*, 2023.

[19] A. Vaswani, N. Shazeer, N. Parmar, et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[20] C. Raffel, N. Shazeer, A. Roberts, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 2020.

[21] T. Brown, B. Mann, N. Ryder, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

[22] F. Petroni, T. Rocktäschel, S. Riedel, et al. Language models as knowledge bases? In *Proceedings of EMNLP*, 2019.

[23] A. Asai, J. Seo, H. Hajishirzi. A survey on retrieval-augmented language models. *arXiv preprint arXiv:2309.05653*, 2023.

[24] L. Gao, K. Ding, et al. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2212.14024*, 2022.

[25] B. Alkhamissi, A. Ray, et al. A review of hallucination in natural language generation. *ACM Computing Surveys*, 2022.

[26] R. Zellers, A. Holtzman, et al. FEVER: A large-scale dataset for fact extraction and verification. In *NAACL*, 2018.

[27] S. Lin, J. Hilton, O. Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *NeurIPS*, 2021.

[28] Y. Feng, H. Wang. A comprehensive survey on text embeddings. *ACM Transactions on Knowledge Discovery from Data*, 2022.

[29] K. Lee, M. Chang, K. Lee, J. Seo. Latent retrieval for weakly supervised open-domain question answering. In *ACL*, 2019.

[30] V. Karpukhin, B. Ogurtsov, S. Riedel, et al. Dense Passage Retrieval for open-domain question answering. In *EMNLP*, 2020.