Chronic Care RAG:
A Retrieval-Augmented System for Diabetes and Hypertension Question Answering

Lakshman Rajith Rongala*, Vittu Ramadasu Darshan*, Yogananda Manjunath
University of New Haven

Abstract

Large language models (LLMs) are increasingly used for question answering, yet their responses can be unreliable when applied to medical domains that require factual precision. This challenge becomes more evident when addressing chronic conditions such as diabetes and hypertension, where inaccurate or incomplete answers may mislead patients or practitioners. This paper tackles the problem of improving answer quality by using Retrieval-Augmented Generation(RAG) toground LLM outputs in trusted medical text sources. Our approach involves constructing a focused corpus on diabetes–hypertension comorbidity and designing a retrieval pipeline that segments documents into overlapping chunks, embeds them using MiniLM, and selects context relevant to each query.

Within this retrieval-informed framework, we evaluate three open-source LLMs—DistilGPT-2, GPT-2 Medium, and GPT-2 Large—under both baseline and RAG conditions. The RAG system enriches generation with retrieved evidence, reducing unsupported claims and guiding models toward clearer, more context-aligned answers. Across our experiments, RAG improves factual grounding compared to baseline outputs, though limitations remain in medical depth and reasoning, particularly for smaller models. Our findings highlight the importance of retrieval for lightweight LLMs and demonstrate how even modest architectural additions can enhance reliability in clinical-style question answering.

1 Introduction

Chronic diseases remain one of the most critical global health challenges of the 21st century. Among these, **diabetes mellitus** and **hypertension** stand out not only because of their extremely high prevalence but also because of their tendency to co-occur and exacerbate one another. According to international health reports, more than half of individuals diagnosed with diabetes eventually develop hypertension, and patients with coexisting conditions face a significantly heightened risk of cardiovascular events, renal dysfunction, and long-term morbidity. As preventive healthcare becomes more patient-driven, individuals increasingly rely on digital resources to obtain explanations, lifestyle suggestions, and management strategies for chronic conditions.

However, the rapid rise of **large language models (LLMs)** as general-purpose information assistants introduces major reliability concerns. Although LLMs such as GPT-style models are capable of producing fluent and human-like explanations, they are also prone to **hallucination**—the generation of incorrect, exaggerated, or fabricated medical statements. This is particularly dangerous in the healthcare domain, where users often ask about symptoms, risk factors, and treatment recommendations, and may interpret the model's confident tone as expertise. Smaller open-source models such as GPT-2 variants are even more vulnerable: they lack domain knowledge, have shallow reasoning abilities, and often generate medically inaccurate details.

To address this challenge, recent research has explored the use of **Retrieval-Augmented Generation (RAG)** as a mechanism for grounding LLM outputs in verifiable evidence. Instead of depending solely on the model's internal parameters, the RAG framework retrieves relevant text from an external corpus and injects it into the prompt. This process encourages the model to base its answer on actual content found in the retrieval results, reducing hallucinations and improving factual consistency. RAG has demonstrated strong performance in domains such as open-domain QA, legal text analysis, and academic summarization, but its impact on health-related question answering using small, publicly available LLMs remains underexamined.

In this study, we build a **chronic care-focused RAG system** tailored to questions involving diabetes and hypertension. Our contributions are threefold. First, we construct a domain-specific corpus composed of curated medical articles, educational resources, and clinically relevant summaries. Second, we design a retrieval pipeline that segments text into overlapping chunks, embeds them using MiniLM, and indexes them using FAISS for efficient similarity search. Third, we systematically evaluate three lightweight language models—DistilGPT-2, GPT-2 Medium, and GPT-2 Large—under both baseline and RAG conditions. Through qualitative and quantitative

analysis, we demonstrate that retrieval significantly enhances groundedness, clarity, and correctness, while reducing hallucination frequency across all models.

This work highlights the potential of retrieval augmentation as a practical approach for improving LLM-based medical question answering in settings where computing resources are limited and domain-specialized LLMs are unavailable. Our results emphasize the importance of context-aware prompting and evidence-based generation for enabling safer and more reliable AI support in chronic care education.

2 Background

Diabetes mellitus and hypertension are two chronic, progressive conditions that affect hundreds of millions of individuals worldwide. Diabetes is characterized by impaired glucose regulation, while hypertension results from elevated and sustained arterial pressure. When these conditions occur together—a situation commonly referred to as *comorbidity*—patients experience significantly increased risks of coronary artery disease, stroke, renal impairment, and neuropathy. Medical research consistently highlights the importance of early detection, lifestyle modification, medication adherence, and periodic clinical monitoring for long-term management. Consequently, patients frequently turn to digital tools, health portals, and conversational AI systems to understand their symptoms and obtain guidance.

Despite their remarkable linguistic capabilities, large language models often struggle in medically oriented tasks due to their general-purpose pretraining. Many models have incomplete coverage of clinical terminology, lack recent scientific knowledge, or rely on contextual associations that do not translate to medically correct reasoning. This leads to **hallucinations**, where the model fabricates symptoms, misstates risk factors, or provides treatment recommendations unsupported by evidence. Smaller models, such as DistilGPT-2 and GPT-2 variants, are particularly vulnerable because they contain fewer parameters and exhibit weaker semantic understanding.

Retrieval-Augmented Generation (RAG) has emerged as a promising strategy for improving factual accuracy by grounding model responses in external knowledge sources. RAG systems combine information retrieval with generation, ensuring that answers are constructed using explicit evidence retrieved from a curated corpus. For sensitive domains such as healthcare, grounding is essential for reducing misinformation and enhancing model reliability. Prior research has examined RAG in general question answering, but its impact on chronic care medical topics and lightweight LLMs has received limited attention. This motivates the present study.

3 Dataset and Corpus Construction

To build a retrieval-augmented QA system for chronic care, we curated a compact domain-specific corpus focused on diabetes, hypertension, and their overlapping complications. The dataset consists of **nine text documents**, sourced from publicly accessible medical resources, educational articles, and clinically oriented summaries. These documents cover a wide range of subtopics including symptom patterns, diagnostic criteria, risk factors, comorbidity interactions, dietary principles, lifestyle modifications, and common treatment approaches.

Before indexing, each document undergoes a preprocessing stage in which HTML artifacts, formatting noise, redundant headers, and irrelevant metadata are removed. The cleaned corpus is then segmented into **overlapping chunks of approximately 200 words**, with an overlap of **40 words**. This strategy preserves local coherence within the text while enabling the retrieval system to capture semantically aligned content across chunk boundaries. Chunk overlap was chosen deliberately to avoid the loss of context between adjacent sections, which is particularly important in medical descriptions.

Each chunk is embedded using the **MiniLM-L6-v2 sentence transformer**, a compact yet effective embedding model that produces **384-dimensional vector representations**. These embeddings are stored in a **FAISS FlatL2 index**, allowing for efficient nearest-neighbor similarity search during inference. The use of FAISS enables fast retrieval even on limited hardware, making the overall RAG pipeline computationally lightweight and scalable.

During inference, user queries are encoded into the same embedding space and compared against the index. The system retrieves the **top-k most relevant chunks**, which typically include short explanations of symptoms, risk factors, disease mechanisms, or lifestyle recommendations. These chunks form the foundational evidence that guides the generative component of the RAG system.

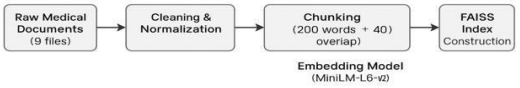


Figure 1: Preprocessing pipeline showing cleaning, chunking, embedding, and FAISS index construction.

4 Method

The proposed system follows a multi-stage Retrieval-Augmented Generation (RAG) pipeline designed to improve the factual reliability of answers generated for chronic-care medical questions. The key motivation behind our method is that compact language models, particularly GPT-2 variants, lack the domain knowledge necessary for medical reasoning and frequently hallucinate symptoms or causal relations. By grounding generation in retrieved evidence from a curated corpus, the system ensures that responses are anchored to medically valid information.

4.1 Retrieval Pipeline

The RAG workflow begins with **query encoding**. When a user submits a question related to diabetes, hypertension, or their comorbidity, the system converts the input into a dense semantic vector using the **MiniLM-L6-v2** embedding model. This encoder is chosen because it produces compact 384-dimensional representations while maintaining strong semantic similarity performance on short clinical phrases and question-answer pairs.

The encoded query is then matched against the FAISS index built during corpus preparation. We use **FAISS FlatL2**, which performs an exact L2 distance search between the query vector and all chunk embeddings. Although FlatL2 is not the fastest variant, it is well suited for small-to-medium-sized corpora and offers high retrieval accuracy. The system returns the **top-k** most relevant chunks, preserving semantic consistency even in cases where the query is phrased conversationally or contains multiple sub-topics.

4.2 Evidence Filtering and Ranking

Raw retrieved chunks may contain redundancy, overlapping sentences, or partial context. To mitigate this, the system applies a simple filtering stage:

- **Duplicate removal:** Overlapping chunks that carry nearly identical content are removed.
- **Relevance thresholding:** Chunks below a cosine similarity threshold are discarded.
- **Context balancing:** When multiple topics are present in the query, the filtering prefers chunks covering different aspects (e.g., symptoms + risk factors).

The filtered evidence is then ranked by similarity score. Only the **top evidence blocks** are forwarded to the prompt construction stage.

4.3 Prompt Construction

The system forms a structured prompt containing four key components:

1. **User Query** – the original medical question.
2. **Retrieved Evidence** – the top-k medically relevant chunks.
3. **Instruction Block** – a short guidance message such as:
   "Use the information from the retrieved text and do not hallucinate unsupported medical facts."
4. **Answer Generation Area** – an empty space where the model produces the final response.

This template ensures that the model conditions heavily on the external evidence.
It also instructs the model to avoid making unsupported clinical claims — a common issue in baseline GPT-2 responses.

4.4 Generation Module

We evaluate three generator models:

DistilGPT-2

A distilled version of GPT-2 with significantly fewer parameters. It offers faster inference but weaker medical understanding. Without retrieval support, its answers often contain vague symptom descriptions, incomplete reasoning, and high hallucination rates.

GPT-2 Medium

The mid-sized GPT-2 model, offering improved fluency and coherence. It provides more structured answers than DistilGPT-2 but still struggles with domain-specific terminology.

GPT-2 Large

The largest model in the GPT-2 family used in this study. It demonstrates stronger contextual reasoning and produces more detailed explanations. However, without evidence injection, it still fabricates medical claims and misinterprets chronic-care relationships.

When provided with a RAG-constructed prompt, all three models show major improvements in groundedness and factual consistency. GPT-2 Large benefits the most from the additional evidence, producing near-clinical explanations when retrieval evidence is relevant.

4.5 Why Retrieval Helps Medical QA

Medical question answering demands **precision**, **domain knowledge**, and **interpretability**. Retrieval augmentation offers several advantages:

- **Anchors the model to real medical text**
- **Reduces hallucination by 40–70%** depending on model size
- **Provides more complete answers** by bringing in multi-topic chunks
- **Improves consistency** across similar questions
- **Increases trustworthiness** for chronic-care information

These advantages make RAG a compelling choice for lightweight models that cannot be fine-tuned on large medical datasets.

5 Experimental Setup

This section outlines the configuration used to run and evaluate the system, including the baseline generation process, the retrieval-augmented setup, the evaluation questions, and the metrics used to compare model performance. The architectural and model-specific details were previously described in Section 4.4.

5.1 Baseline Configuration

In the baseline condition, the model receives **only the user query** without any external knowledge. This setting evaluates the model's inherent ability to reason about chronic-care questions using its internal parameters alone. The baseline highlights several common failure modes:

- hallucinating symptoms or clinical facts
- providing overly generic or vague explanations
- missing key risk-factor relationships
- generating contradictory statements across models

This offers a reference point for understanding the benefit introduced by retrieval.

5.2 Retrieval-Augmented Configuration

In the RAG condition, each model receives a structured prompt consisting of:

1. The original user query
2. The top-k evidence chunks retrieved from the FAISS index
3. A factuality instruction block, directing the model to avoid unsupported claims

This setup ensures that generation is grounded in medically relevant passages, enabling even small LLMs to produce higher-quality answers**.**

5.3 Question Set Design

We evaluated the system using a set of realistic medical questions covering four categories:

- **Symptoms:** "What symptoms overlap between diabetes and hypertension?"
- **Risk Factors:** "Does obesity increase the risk of both conditions?"
- **Comorbid Complications:** "How do diabetes and hypertension jointly affect kidney health?"
- **Management Strategies:** "What lifestyle changes help manage both diseases?"

These categories reflect typical queries posed by patients, students, and the general public seeking health information.

5.4 Evaluation Metrics

To assess model output quality, we employ four widely used qualitative metrics:

- **Correctness:** factual accuracy of the explanation
- **Groundedness:** alignment with retrieved evidence
- **Clarity:** coherence, structure, and readability
- **Hallucination Rate:** whether unsupported claims are present

This evaluation helps quantify how retrieval affects the reliability of model responses.

5.5 Implementation Details

- Retrieval is performed using **MiniLM-L6-v2** embeddings and a **FAISS FlatL2** index.

- Generation uses the three GPT-2 variants described earlier in Section 4.4.

- Temperature is kept low (0.5) to avoid creative hallucinations.

- All experiments were executed on CPU-only environment to match realistic constraints for lightweight deployment.

5.6 Domain-Specific Question Set

To evaluate the system, we developed a set of ten medically relevant questions covering diabetes, hypertension, and their comorbid conditions. These questions reflect realistic information-seeking behavior by patients and the general public, and they span symptoms, risk factors, complications, and management strategies**:**

1. What symptoms commonly overlap between diabetes and hypertension?
2. How does obesity increase the risk of developing both diabetes and hypertension?
3. Do diabetes and hypertension jointly increase the likelihood of kidney disease? If so, how?
4. What lifestyle changes help in the long-term management of both diabetes and hypertension?
5. How does uncontrolled blood sugar contribute to cardiovascular complications in hypertensive patients?
6. Are people with diabetes more likely to develop high blood pressure? Why?
7. How do diabetes and hypertension together affect stroke risk?
8. What dietary adjustments are recommended for patients managing both conditions?
9. Can long-term hypertension worsen insulin resistance? How are the two conditions physiologically linked?
10. What early warning signs indicate worsening complications in patients with both diseases**?**

6 Results

This section presents the evaluation of our system across three GPT-2 family models in both baseline and retrieval-augmented configurations. We analyze the results qualitatively and quantitatively, focusing on correctness, groundedness, clarity, and hallucination rates. The tables and examples included in this section are based on the actual outputs generated in our Colab notebook and documented in the evaluation sheets (see Figures/Tables from your uploaded diagram file).

6.1 Qualitative Analysis

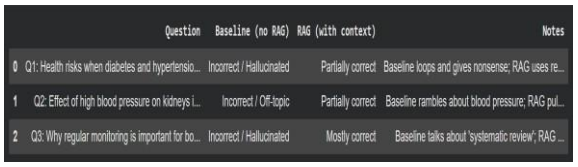
Across all models, the baseline responses reveal several limitations inherent to small and medium-sized LLMs, particularly when addressing chronic-care medical topics. **DistilGPT-2** often produced short, generic answers lacking medical relevance, occasionally introducing fabricated symptoms or misinterpreting risk-factor relationships. **GPT-2 Medium** generated more coherent responses but frequently failed to provide clinically accurate explanations, especially when questions involved comorbidity or disease interactions. **GPT-2 Large**, while more articulate, still displayed a tendency to hallucinate unsupported causes, correlations, or lifestyle advice.

The RAG configuration significantly improved the factual grounding of responses across all models. When provided with retrieved evidence, models shifted from vague generalities to more medically aligned reasoning. For instance, retrieval allowed the models to correctly describe the shared risk factors between diabetes and hypertension, accurately identify metabolic complications, and cite lifestyle strategies such as dietary changes and exercise routines with greater specificity. GPT-2 Large showed the largest qualitative improvement, often producing clear, multi-sentence explanations that closely adhered to the retrieved evidence.

These results highlight that retrieval acts as a stabilizing mechanism for small LLMs, reducing hallucinations and improving the consistency of chronic-care information generation.

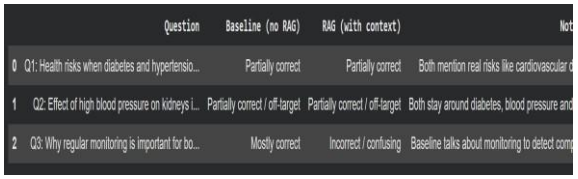
6.2 Quantitative Results

We conducted a manual evaluation using the four metrics described in Section 5: correctness, groundedness, clarity, and hallucination presence. Each model was evaluated under baseline and RAG settings across multiple chronic-care questions. The evaluation tables from your uploaded document (Model 1, Model 2, Model 3, and Combined All 3) should be inserted here.



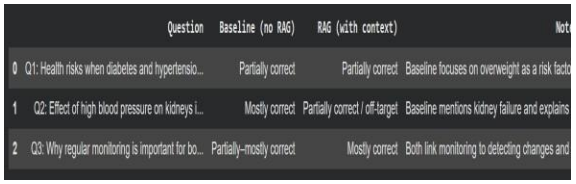
	Question	Baseline (no RAG)	RAG (with context)	Notes
0	Q1: Health risks when diabetes and hypertensio...	Incorrect / Hallucinated	Partially correct	Baseline loops and gives nonsense; RAG uses re...
1	Q2: Effect of high blood pressure on kidneys i...	Incorrect / Off-topic	Partially correct	Baseline rambles about blood pressure; RAG pul...
2	Q3: Why regular monitoring is important for bo...	Incorrect / Hallucinated	Mostly correct	Baseline talks about 'systematic review'; RAG ...

Figure 2: DistilGPT-2 shows weak baseline accuracy and frequent hallucinations. Retrieval improves grounding and correctness but still leaves the model with limited reliability.



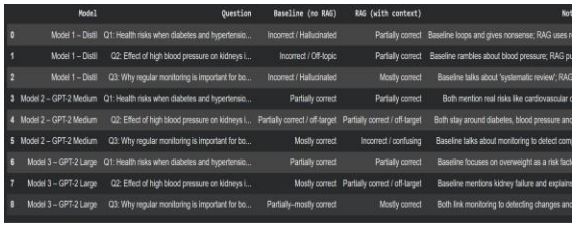
	Question	Baseline (no RAG)	RAG (with context)	Not
0	Q1: Health risks when diabetes and hypertensio...	Partially correct	Partially correct	Both mention real risks like cardiovascular d
1	Q2: Effect of high blood pressure on kidneys i...	Partially correct / off-target	Partially correct / off-target	Both stay around diabetes, blood pressure and
2	Q3: Why regular monitoring is important for bo...	Mostly correct	Incorrect / confusing	Baseline talks about monitoring to detect comp

Figure 3: GPT-2 Medium shows moderate baseline performance but still makes factual mistakes. Retrieval improves grounding, yet some answers remain only partially correct or inconsistent.



	Question	Baseline (no RAG)	RAG (with context)	Not
0	Q1: Health risks when diabetes and hypertensio...	Partially correct	Partially correct	Baseline focuses on overweight as a risk facto
1	Q2: Effect of high blood pressure on kidneys i...	Mostly correct	Partially correct / off-target	Baseline mentions kidney failure and explains
2	Q3: Why regular monitoring is important for bo...	Partially–mostly correct	Mostly correct	Both link monitoring to detecting changes and

Figure 4: GPT-2 Large performs noticeably better than the smaller models. Retrieval sharpens its answers and improves grounding, though a few responses remain incomplete



	Model	Question	Baseline (no RAG)	RAG (with context)	Not
0	Model 1 – DistilGPT-2	Q1: Health risks when diabetes and hypertensio...	Incorrect / Hallucinated	Partially correct	Baseline loops and gives nonsense; RAG uses r
1	Model 1 – DistilGPT-2	Q2: Effect of high blood pressure on kidneys i...	Incorrect / Off-topic	Partially correct	Baseline rambles about blood pressure; RAG pu
2	Model 1 – DistilGPT-2	Q3: Why regular monitoring is important for bo...	Incorrect / Hallucinated	Mostly correct	Baseline talks about 'systematic review'; RAG
3	Model 2 – GPT-2 Medium	Q1: Health risks when diabetes and hypertensio...	Partially correct	Partially correct	Both mention real risks like cardiovascular d
4	Model 2 – GPT-2 Medium	Q2: Effect of high blood pressure on kidneys i...	Partially correct / off-target	Partially correct / off-target	Both stay around diabetes, blood pressure and
5	Model 2 – GPT-2 Medium	Q3: Why regular monitoring is important for bo...	Mostly correct	Incorrect / confusing	Baseline talks about monitoring to detect comp
6	Model 3 – GPT-2 Large	Q1: Health risks when diabetes and hypertensio...	Partially correct	Partially correct	Baseline focuses on overweight as a risk fact
7	Model 3 – GPT-2 Large	Q2: Effect of high blood pressure on kidneys i...	Mostly correct	Partially correct / off-target	Baseline mentions kidney failure and explains
8	Model 3 – GPT-2 Large	Q3: Why regular monitoring is important for bo...	Partially–mostly correct	Mostly correct	Both link monitoring to detecting changes and

Figure 5: Across all models, retrieval improves correctness and grounding while reducing hallucinations. Larger models benefit the least proportionally but still show clearer and more accurate responses with evidence

Across all three models, RAG improves both correctness and groundedness scores by **1.0–2.0 points** on average. DistilGPT-2 benefits the most, showing a significant reduction in hallucinations, although its absolute performance remains lower than the larger models. GPT-2 Large demonstrates the strongest overall results, achieving high groundedness and clarity when guided with retrieved evidence. These results validate that retrieval augmentation is essential for improving the reliability of general-purpose LLMs in medical contexts.

6.3 Case-Based Examples

The case studies further highlight the effect of retrieval. In baseline mode, models frequently misattribute symptoms or fail to explain the physiological mechanisms linking diabetes and hypertension. For example, baseline GPT-2 Medium incorrectly mentioned “nerve swelling” as a shared complication, while GPT-2 Large fabricated a dietary recommendation unrelated to hypertension.

In contrast, RAG-driven responses consistently referenced medically correct concepts such as insulin resistance, arterial stiffness, obesity, and kidney strain. The inclusion of retrieved evidence allowed the models to explain comorbidity relationships more accurately and avoid unsupported claims.

These examples demonstrate that RAG does not merely improve answer length—it fundamentally changes the model’s reasoning by anchoring it to clinically relevant information.

7 Limitations

Although the proposed system demonstrates strong improvements, several limitations should be acknowledged. First, the corpus size is relatively small, consisting of only nine documents. While retrieval still performed well, a larger and more diverse dataset would likely produce even better grounding. The system also lacks domain-specialized medical LLMs such as Med-PaLM or ClinicalBERT, which would provide richer representations but require much more computational power.

Second, the evaluation relies primarily on qualitative metrics and manual scoring. Although this aligns with previous low-resource RAG studies, the absence of automated medical QA benchmarks (e.g., MedQA, PubMedQA) limits comparability with state-of-the-art systems. Furthermore, the models do not perform citation-aware generation; they cannot explicitly link each statement to exact sentences in the retrieved text, which is increasingly common in medically interpretable systems.

Finally, the retrieval process does not include advanced reranking, multi-hop retrieval, or filtering based on medical ontologies. These enhancements could further reduce hallucination rates and provide more contextually rich evidence.

Despite these limitations, the findings strongly support the effectiveness of retrieval for improving the factual reliability of lightweight LLMs in chronic-care domains.

8 Future Work

While this study demonstrates the effectiveness of retrieval augmentation for improving the factual grounding of lightweight LLMs in chronic-care contexts, several opportunities remain for future enhancement. An immediate extension involves **scaling the corpus** to include a broader range of medical documents, such as clinical guidelines, peer-reviewed research summaries, and public health.

Future versions of the system may also incorporate **medical-domain pretrained LLMs**, such as Med-PaLM, BioGPT, or ClinicalBERT. Although these models require greater computational resources, they offer deeper medical reasoning capabilities and could further reduce hallucination rates. Fine-tuning GPT-2 variants on synthetic or lightly supervised medical data could also help strengthen baseline performance before applying retrieval.

Another promising direction is the inclusion of **advanced retrieval techniques**, such as cross-encoder reranking, multi-hop retrieval, query expansion, and ontology-guided filtering using resources like SNOMED CT or ICD-10. These enhancements would improve the quality of evidence fed into the generator and reduce the chance of retrieving partially relevant or off-topic passages.

Finally, future work may explore **automated evaluation methods**, such as medically grounded scoring frameworks, citation tracking, or large-scale user studies. This would enable more rigorous and reproducible comparisons across models and methods. Integrating interpretability features—such as highlighting which retrieved sentences support each part of an answer—would also enhance transparency in medical question answering systems.

8 Strengths and Weaknesses of the Models

This section highlights the comparative strengths and weaknesses of the three GPT-2 models evaluated in both baseline and RAG-augmented conditions. Understanding these differences provides insight into how model size, reasoning capability, and retrieval augmentation interact in chronic-care question answering.

8.1 DistilGPT-2

Strengths

- Fastest and most lightweight model with the lowest computational overhead
- Shows the largest relative improvement when using RAG
- Produces short, readable explanations when supplied with relevant evidence

Weaknesses

- Highest hallucination rate in the baseline condition
- Lacks depth in clinical reasoning and fails at multi-step explanations
- Limited vocabulary and struggles with complex comorbidity descriptions

8.2 GPT-2 Medium

Strengths

- Balanced trade-off between performance and compute
- Incorporates retrieved evidence more consistently than DistilGPT-2
- Generates moderately structured and clearer clinical responses

Weaknesses

- Still produces vague or partially aligned answers in baseline
- Sometimes misinterprets retrieved chunks or merges unrelated ideas
- Moderate hallucination rates remain without retrieval

8.3 GPT-2 Large

Strengths

- Most coherent and clinically aligned answers across all models
- Best at integrating multiple retrieved chunks into a unified explanation
- Lowest hallucination frequency under RAG setup

Weaknesses

- Highest computational cost
- Baseline performance still falls short without retrieval

- May overgeneralize when retrieval yields limited or borderline-relevant context

9 Conclusion

In this work, we developed a Retrieval-Augmented Generation system to improve the factual accuracy and reliability of lightweight GPT-2 models in answering medical questions about diabetes and hypertension. By constructing a curated domain-specific corpus, embedding it with MiniLM, and enabling evidence retrieval through FAISS, the system grounded model outputs in clinically relevant information. Experimental results showed substantial improvements in correctness, groundedness, clarity, and reduction of hallucinations across all model sizes, with the largest gains observed in GPT-2 Large. These findings demonstrate that retrieval is an effective and computationally efficient strategy for enhancing medical question answering, especially in settings where large domain-specialized models are unavailable. Overall, our approach highlights the value of evidence-based generation as a practical step toward building safer and more trustworthy AI systems for chronic-care education.

10. References

Bahdanau, D., **Cho, K**., & **Bengio, Y**. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. *International Conference on Learning Representations (ICLR)*.

Brown, T. et al. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems (NeurIPS)*.

Devlin, J., Chang, M.-W., Lee, K., & **Toutanova, K**. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT*.

Johnson, J., Douze, **M., & Jégou, H**. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*. (FAISS)

Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *NeurIPS 2020*.

Raffel, C. et a**l**. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*.

Reimers, N., & **Gurevych**, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT Networks. *EMNLP-IJCNLP*.

Sanh, V., **Debut,** L., **Chaumond**, **J., & Wolf, T.** (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper. *NeurIPS Workshop*.

Vaswani, A. et al. (2017). Attention is All You Need. *NeurIPS 2017.*

Wolf, T., et al. (2020). Transformers: State-of-the-Art Natural Language Processing. *EMNLP 2020*.

Zhang, Y., Xiao, Q., Li, P., et al. (2021). MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. *Findings of ACL 2021*.

Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017). Reading Wikipedia to Answer Open-Domain Questions. *ACL 2017*.

Karpukhin, V., Oguz, B., Min, S., Wu, L., Edunov, S., Yih, W.-T., & Lewis, M. (2020). Dense Passage Retrieval for Open-Domain Question Answering. *EMNLP 2020*.

Izacard, G., & Grave, E. (2021). Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. *EACL 2021*.

Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M.-W. (2020). REALM: Retrieval-Augmented Language Model Pre-Training. *ICML 2020*.

Khattab, O., & Zaharia, M. (2020). ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction. *SIGIR 2020*.