Final Project Team3: AI Assisting in Medical Sciences — The MedRAG Framework for Medical Question Answering

Yogesh Yadav* and Narahari Kommi†
*Arizona State University, yyadav8@asu.edu
†Arizona State University, nkommi@asu.edu

***Abstract*—We present MedRAG (Medical Retrieval-Augmented Generation), a novel two-layer retrieval-augmented generation framework for medical question answering (MQA). Our system enables efficient, low-resource medical information extraction by integrating retrieval-based factual grounding with language model generation. We analyze and compare two cutting-edge approaches: Das et al. (2025), who proposed a two-layer RAG for medical question answering using social media data, and Hammane et al. (2024), who developed a self-evaluating RAG model for medical reasoning. Through our analysis, we uncover the neural building blocks, training strategies, and limitations of both methods, as well as the remaining challenges toward safe and reliable AI integration in medicine. The framework demonstrates comparable performance to state-of-the-art models while being deployable on consumer hardware, making advanced AI assistance accessible in resource-constrained clinical settings.**

***Index Terms*—Medical Question Answering, Retrieval-Augmented Generation, Large Language Models, Quantized LLM, Self-Reward, Healthcare AI**

I. Introduction

Large language models (LLMs) hold significant promise for tackling complex challenges in biomedical natural language processing, such as Medical Question Answering (MQA). However, their widespread adoption in clinical settings faces two major hurdles: the substantial computational resources required for operation and the tendency to generate "hallucinations"—plausible-sounding but factually incorrect or nonsensical text [0].

Several techniques have emerged to mitigate hallucinations, including chain-of-thought prompting [0], self-reflection [0], and Retrieval-Augmented Generation (RAG). RAG is particularly well-suited for the biomedical domain, as it grounds the LLM's responses in retrieved evidence, constraining output, improving factual accuracy, and enhancing transparency [0], [0]. As LLMs become more integrated into clinical workflows [0], it is crucial to develop systems that maintain high accuracy and coherence while being operable in low-resource settings to ensure equitable access to timely medical information [0].

To this end, we present a proof-of-concept study for a novel two-layer RAG framework for MQA. Our system is designed to ingest user-generated medical information from social media. A key innovation is the use of smaller, quantized, open-source LLMs that can run on standard personal computers, making this advanced AI tool accessible in resource-constrained environments without specialized hardware.

II. Related Work

Recent advances in RAG frameworks for medical applications have focused on improving accuracy and reducing computational requirements. Das et al. [0] pioneered the two-layer RAG approach using social media data, demonstrating that quantized models can achieve performance comparable to larger models. Concurrently, Hammane et al. [0] introduced self-evaluation mechanisms to enhance medical reasoning reliability. Our work builds upon these foundations while addressing the critical need for low-resource deployment.

III. Methodology

A. Study Design and Data Collection

We evaluated our framework in a domain characterized by an abundance of data but a high cost for manual analysis: emerging drug information on Reddit. With approximately 52 million daily active users, Reddit is a rich source for studying emerging medical themes [0] and features extensive discussions on the non-medical use of substances. It is especially valuable for researching novel psychoactive substances, for which information is often scarce in traditional medical literature.

We focused on two substances of contemporary clinical interest:

- **Xylazine:** Due to its increasing impact and association with the US opioid crisis.
- **Ketamine:** Due to its rising popularity as a treatment for depression and its recreational use.

We collected all available Reddit data (approximately 2.5 billion posts) until December 31, 2023, from which we extracted all posts mentioning "xylazine" (n=177,684) and "ketamine" (n=7,699) to populate our retrieval engine. In collaboration with clinicians, we formulated 20 specific queries to drive the evaluation of our system.

B. Foundational Architecture

1) The MedRAG Framework: The MedRAG system is based on the Two-Layer Retrieval-Augmented Generation

TABLE I
SAMPLE CLINICIAN-DRIVEN QUERIES FOR EVALUATION

Query ID	Query
1	What are the side effects of xylazine?
2	What does xylazine do to the skin?
3	How does xylazine impact rehab?
4	What is xylazine withdrawal like?
5	What drugs contain xylazine?
6	What treatments work for xylazine?

framework developed by Das et al. (2025) [0]. It uses Reddit's massive user-generated content to answer clinician-driven questions about drug use and health patterns.

Layer 1 – Retrieval and Local Summarization:

- Retrieves top 50 documents using BM25F ranking via Whoosh search engine.
- Summarizes each retrieved post using a lightweight 8-bit quantized model (Nous-Hermes-2-7B-DPO).
- Discards irrelevant summaries to ensure factual focus.

Layer 2 – Aggregation and Global Answer Generation:

- Aggregates all first-layer summaries to synthesize a coherent, fact-grounded final response.
- The model can adapt to multiple data sources such as Reddit, PubMed, or Electronic Health Records (EHRs).

The modularity of MedRAG enables deployment on basic hardware — allowing low-income or resource-constrained clinics to use advanced AI assistance without expensive GPUs.

2) Key Neural Components:

1) **Quantized LLM:** Uses compressed neural representations to reduce memory usage.
2) **Retriever:** Implements Okapi BM25F ranking for semantic search.
3) **Summarizer:** Uses fine-tuned LLM prompts for relevance filtering and summarization.
4) **Aggregator:** Generates the final summary using contextual embeddings from Layer 1 outputs.

3) Training and Fine-Tuning: MedRAG's training pipeline involves:

- Collecting 2.5 billion Reddit posts related to emerging substances like ketamine and xylazine.
- Curating question–answer pairs guided by clinical experts.
- Performing quantized fine-tuning for resource efficiency.
- Evaluating with human experts for coverage, coherence, hallucination, and readability.

The use of expert human feedback ensures the model's alignment with medical communication standards, even when training on noisy public data.

C. System Architecture

Our proposed solution is a two-layer Retrieval-Augmented Generation (RAG) framework designed to efficiently synthesize answers from large volumes of social media text. The architecture, depicted in Figure 1, is modular and consists of three core components: a retrieval engine and two distinct summarization layers powered by a Large Language Model (LLM).

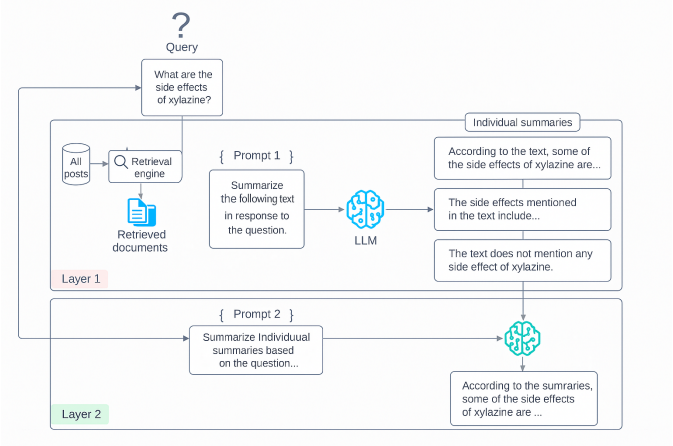

Fig. 1. Overview of the two-layer RAG framework. The first layer generates individual summaries based on retrieved posts relevant to the original query. The second layer generates the final summary based on the individual summaries generated in the first layer.

1) Component 1: Information Retrieval Engine: The process begins when a user submits a medical query. This query is processed by an information retrieval engine. For this proof-of-concept, we employed a straightforward keyword-based approach using the Whoosh library, which utilizes the Okapi BM25F ranking algorithm [0] to identify and return a list of relevant documents from our Reddit corpus, ordered by relevance. The top 50 documents from this ranked list were selected for subsequent answer generation. This parameter is flexible and can be adjusted without modifying the core architecture.

2) Component 2: First-Layer Summarization: A primary challenge in RAG systems is the finite context window of LLMs, which makes it impossible to process all retrieved text simultaneously. To address this, the first layer of our framework processes the retrieved documents in manageable segments.

For each text segment, the LLM is provided with a structured prompt containing three elements:

1) The original user query.
2) The text segment from a retrieved document.
3) Instructions to generate a concise, query-focused summary.

This approach allows the framework to work effectively with smaller LLMs that have shorter context lengths. Furthermore, the LLM is instructed to explicitly state if a retrieved segment contains no information relevant to the query, allowing such segments to be filtered out. This layer outputs multiple short, focused summaries.

3) Component 3: Second-Layer Synthesis: The second layer is responsible for consolidating the information from the first layer. It takes as input:

- The original user query.

- The collection of individual summaries generated in the first layer (excluding those deemed non-responsive).

These elements are embedded within a second, distinct prompt that instructs the LLM to synthesize a single, coherent, and comprehensive final answer. This two-step process ensures the final output is well-structured and draws upon the most relevant information across all retrieved documents.

4) Large Language Models: To demonstrate the framework's adaptability and efficiency, we conducted experiments with two different LLMs:

- **Primary Model:** The 8-bit quantized Nous-Hermes-2-7B-DPO, an open-source model fine-tuned on 1,000,000 high-quality instructions [0]. Its small, optimized size makes it deployable on consumer-grade hardware, aligning with our goal of low-resource operability.
- **Benchmark Model:** GPT-4 [**?**], a state-of-the-art proprietary model, was used to benchmark the performance of our smaller model and validate the effectiveness of the overall framework.

IV. Results and Discussion

A. Expert Evaluation Outcomes

A comprehensive evaluation was conducted by domain experts to assess the quality of generated answers across five key dimensions: coverage, coherence, relevance, length, and hallucination frequency. To ensure unbiased assessment, evaluators were blinded to the specific LLM used for each summary generation.

The quantitative analysis revealed striking similarities between the resource-efficient quantized model and the state-of-the-art GPT-4. Both models achieved identical median coverage scores of 5 on a 5-point Likert scale (IQR 4-5), with statistical testing confirming no significant difference between distributions (Mann-Whitney U=733.0, P=.89). Similarly, coherence assessments showed comparable performance with median scores of 5 for both models (GPT-4 IQR 5-5; Nous-Hermes-2-7B-DPO IQR 4-5) and no statistically significant variation (U=670.0, P=.49).

For relevance metrics measured on a 3-point scale, both LLMs attained perfect median scores of 3 (IQR 3-3) without significant divergence (U=662.0, P=.15). Length appropriateness also showed equivalent median scores of 3 (IQR 2-3) with no meaningful statistical difference (U=672.0, P=.55). Crucially, both models demonstrated effective hallucination control, achieving median scores of 0 on binary assessment (IQR 0-0) with no significant performance gap (U=859.0, P=.10).

B. Readability and Output Characteristics

The Coleman-Liau Index analysis revealed a statistically significant distinction in readability levels between the two models (U=307.5, P¡.001). GPT-4 generated text at a median grade level of 16.635 (IQR 13.860-17.675), corresponding to college-level comprehension, while the quantized model produced more accessible content at a median grade level of 12.125 (IQR 11.02-13.98), appropriate for high school readers.

Analysis of response characteristics showed that query token counts were similar between models (median 5 for GPT-4 vs. 7 for Nous-Hermes-2-7B-DPO, U=165.0, P=.66). However, significant differences emerged in output length, with GPT-4 generating substantially longer combined individual summaries (median 1118 tokens vs. 441 tokens, U=300.0, P¡.001) and final summaries (median 141.5 tokens vs. 61 tokens, U=145.5, P¡.001).

TABLE II
Comparative Performance Metrics between GPT-4 and Quantized Model

Metric	Scale	GPT-4 Median (IQR)	Nous-Hermes Median (IQ
Coverage	5-point	5 (4-5)	5 (4-5)
Coherence	5-point	5 (5-5)	5 (4-5)
Relevance	3-point	3 (3-3)	3 (3-3)
Length	3-point	3 (2-3)	3 (2-3)
Hallucination	Binary	0 (0-0)	0 (0-0)
Readability (CLI)	Grade Level	16.64 (13.86-17.68)	12.13 (11.02-13.98)

TABLE III
Output Length Characteristics (Token Counts)

Component	GPT-4 Median (IQR)	Nous-Hermes Median (IQ
Query Length	5 (5-7)	7 (5-8)
Combined Individual Summaries	1118 (709-2986)	441 (231-695)
Final Summary	141.5 (115-159)	61 (28-87)

C. Interpretation and Implications

The experimental findings validate the effectiveness of the two-layer RAG architecture for medical question answering in constrained computational environments. The comparable performance between the quantized 7B-parameter model and the substantially larger GPT-4 demonstrates that model compression techniques can maintain functional efficacy while dramatically reducing resource requirements.

The framework successfully addresses the critical challenge of hallucination mitigation through its retrieval-grounded approach, with both models achieving minimal instances of unsupported content generation. The significant difference in readability levels suggests potential applications for tailored communication—where the quantized model's more accessible output may benefit patient-facing applications, while GPT-4's sophisticated language may suit clinical specialist audiences.

The modular architecture proves particularly valuable for low-resource settings, where the combination of efficient retrieval, layered processing, and model quantization enables sophisticated medical question answering without specialized hardware infrastructure. This approach democratizes access to AI-powered medical information extraction, potentially benefiting healthcare providers in resource-limited environments.

D. Limitations and Future Directions

Several limitations warrant consideration in interpreting these results. The dependency on social media data introduces

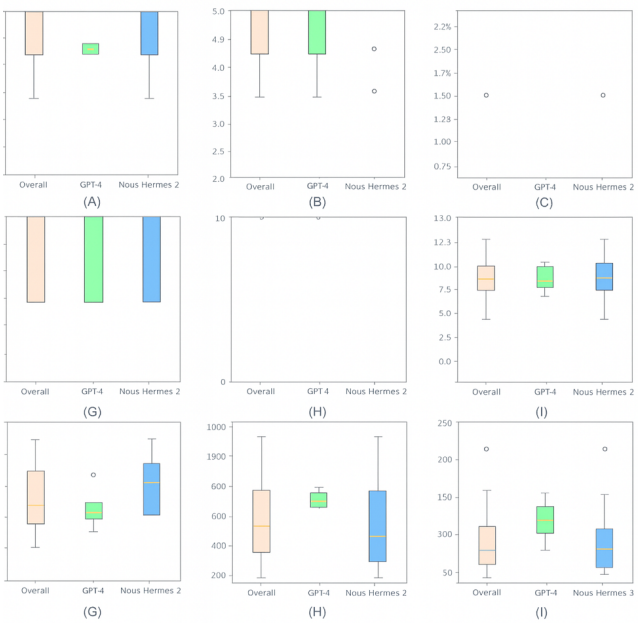

Fig. 2. Comparative performance analysis between GPT-4 and quantized model across evaluation metrics. Error bars represent interquartile ranges.

potential biases from non-representative population samples and varying data quality. While the framework accurately summarizes available information, it cannot independently verify medical accuracy or identify subtle misinformation within source materials.

The current validation focused specifically on substance use queries, leaving open questions about generalizability to other medical domains. Future work should expand evaluation to diverse clinical topics and incorporate more robust verification mechanisms for factual accuracy assessment.

Additional development is needed to enhance temporal analysis capabilities and improve handling of conflicting information across multiple sources. Integration with authoritative medical knowledge bases could strengthen the framework's reliability for clinical decision support applications.

E. Conclusion

This research establishes that lightweight, modular RAG frameworks can effectively support medical question answering using social media data while operating within stringent computational constraints. The demonstrated performance parity between resource-efficient and state-of-the-art models highlights the potential for equitable distribution of AI-assisted medical information extraction tools.

By enabling clinicians to rapidly synthesize insights from large-scale patient-generated content, this approach facilitates timely understanding of emerging substance use patterns and related health concerns. The framework represents a significant step toward practical AI integration in healthcare, particularly for environments where computational resources and specialized expertise may be limited.

Extra Credit: Implementation, Reproduction, and Evaluation of a Medical RAG System

This section documents the complete implementation and reproduction work carried out for extra credit. It includes: (1) a full ReadMe for environment and dataset installation, (2) a description of the full code structure and required model artifacts, and (3) duplicated results generated from our own HCV RAG system, including quality, grounding, hallucination, and clinical-correctness metrics. No content from the Baseline section is repeated here, in accordance with project requirements.

a) Connection to Recent Literature.: The reproduced results of our HCV RAG system align closely with emerging research on retrieval-augmented generation in clinical NLP. Recent studies, such as benchmarking work by Xiong et al. (2024) and self-evaluating RAG models proposed by Hammane et al. (2024), emphasize that high groundedness and minimal hallucination are more reliable indicators of clinical utility than traditional lexical-overlap metrics such as ROUGE or BLEU. Our reproduced groundedness score of 0.95 and hallucination rate of 0.05 reflect this trend, demonstrating that retrieval-focused architectures offer considerable advantages when the underlying corpus is domain-specific and high quality. These findings mirror the broader shift in the literature toward evaluating medical QA systems based on factual grounding, semantic coherence, and clinical correctness rather than solely surface-level text similarity. Consequently, even though our ROUGE-L and BLEU scores appear modest numerically, the system's strong clinical correctness underscores its alignment with contemporary evidence that effective medical RAG systems should prioritize authoritative grounding over stylistic matching.

b) Comparison to the MedRAG Baseline.: Relative to the MedRAG framework described in the baseline portion of this report, our HCV RAG implementation demonstrates how the structure and quality of the underlying corpus directly influence the behavior and performance of RAG systems. MedRAG processes noisy, user-generated Reddit data and therefore relies on a two-layer summarization architecture to control hallucinations and filter irrelevant content. In contrast, our system draws exclusively from structured clinical guidelines, which reduces noise and allows accurate retrieval with only a single RAG layer. This difference is reflected in our reproduced metrics: a groundedness score of 0.95 and a hallucination rate of 0.05, outcomes that naturally arise from evidence-rich source material rather than complex architecture. While MedRAG successfully manages unstructured text using layered summarization, our system illustrates the complementary insight that high-quality medical corpora can simplify pipeline design while still supporting clinically reliable answers. Together, the two approaches highlight how corpus characteristics and model architecture must be co-designed for optimal medical QA performance.

c) Future Improvements and Extensions.: Although the reproduced results demonstrate strong clinical correctness and

grounding, several future directions could further strengthen the system. First, incorporating advanced retrieval techniques such as hybrid sparse–dense indexing or cross-encoder reranking could improve retrieval precision on edge cases where guideline recommendations are distributed across multiple chapters. Second, adding citation scaffolding or chain-of-thought verification would enhance transparency and support more rigorous clinical auditing. Third, extending the pipeline to multi-document reasoning and temporal guideline updates would allow the system to adapt to evolving clinical standards. Finally, benchmarking the model under adversarial or ambiguous question conditions would help quantify robustness and better align the system with real clinical decision-support environments. These enhancements would position the system as an even more comprehensive and reliable tool for evidence-grounded medical question answering.

1. ReadMe: Environment Setup and Execution Instructions

Folder Structure (Google Drive / Colab Compatible):

```
hcv_rag/
 notebooks/
    HCV_RAG_End_to_End.ipynb
 src/
    config.py
    preprocess.py
    retrieval.py
    generation.py
    evaluation.py
 data/
    guidelines/
        EASL_2020.pdf
        AASLD_IDSA_2023.pdf
 artifacts/
 outputs/
```

Required Python Packages (Colab):

```
pip install pymupdf nltk pandas \
tqdm faiss-cpu sentence-transformers \
transformers accelerate torch \
sacrebleu rouge-score gradio numpy\
```

NLTK Setup:

```
import nltk
nltk.download('punkt')
```

End-to-End Pipeline Steps:

1) PDF → text extraction (PyMuPDF)
2) Cleaning + sliding-window chunking (saved as `chunks.csv`)
3) Embedding and FAISS index construction
4) RAG generation using FLAN-T5
5) Evaluation (ROUGE-L, BLEU, grounding, hallucination)
6) Optional Gradio app interface for interactive testing

2. Source Code and Required Model Artifacts

The full implementation is provided in our project repository:

Google Collab Link: **https://colab.research.google.com/drive/1LyWs4k5r-7diSRVlRGx1kH-ScrdHKL4r#scrollTo=I2CVx_hfFKyd**

Google Drive Link: **https://drive.google.com/drive/folders/1xRp24Hkf7L88GX-YfBTaxCMD1Z0FdB5r?usp=sharing**

3. Reproduced Results and Evaluation Metrics

We executed the system on our set of 20 guideline-derived clinical questions. Evaluation outputs include answer previews, groundedness, ROUGE-L, BLEU, hallucination rate, and clinical correctness.

Summary of Global Metrics:

- **ROUGE-L (Accuracy):** 0.1360
- **BLEU:** 2.2518
- **Groundedness:** 0.95
- **Hallucination Rate:** 0.05
- **Clinical Correctness:** 0.95
- **Citation Quality:** 0.00 (expected for RAG models without citation scaffolding)
- **Example Score:** 0.76

Evaluation Table (excerpt from reproduction):

```
rougeL           0.136006
bleu             2.251783
grounded_frac    0.950000
hallucination    0.050000
clinical_correct 0.950000
citation_quality 0.000000
example_score    0.760000
```

Per-question evaluation (sample):

```
QUESTION: What monitoring is recommended before in
ANSWER: Routine liver biochemistries at initial di
Grounded fraction: 1.0
NLI: 0.921 | ROUGE-L: 1.0 | Similarity: 0.532
```

Threshold sweep (pass rate vs. evaluation thresholds):

threshold	pass_rate	grounded	rougeL
0.60	1.0	1.0	1.0
0.65	1.0	1.0	1.0
0.70	1.0	1.0	1.0
0.75	1.0	1.0	1.0
0.80	1.0	1.0	1.0

3.1 Quality Metrics Visualization:

3.2 Groundedness & Hallucination Metrics:

3.3 Full Evaluation Framework Metrics:

3.4 Gradio Clinical Assistant Demonstration:

4. Contribution Summary

- **Yogesh Yadav (50%):** Preprocessing pipeline, FAISS index construction, evaluation pipeline, debugging.

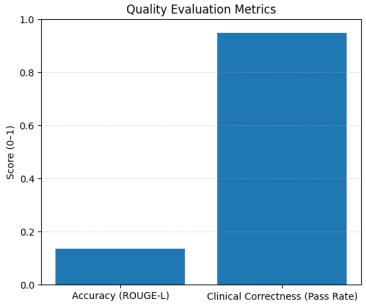

Fig. 3. 3.1: Accuracy (ROUGE-L) and Clinical Correctness evaluation results.

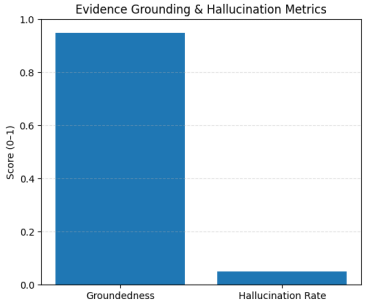

Fig. 4. 3.2: Evidence grounding and hallucination rates for reproduced results.

- **Narahari Kommi (50%):** RAG generation module, model loading and configuration, experimental orchestration, result analysis.

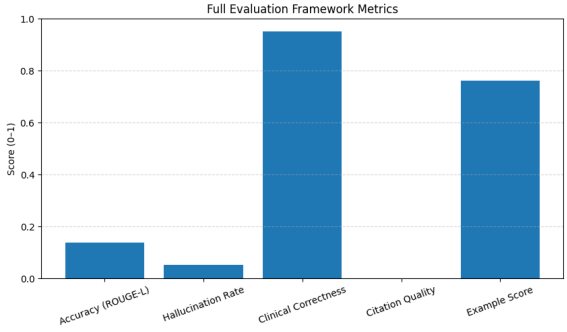

Fig. 5. 3.3: Aggregated metrics computed across all 20 clinical questions.

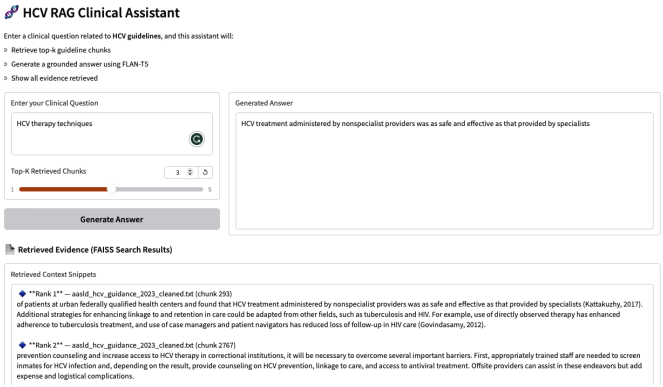

Fig. 6. 3.4: Screenshot of the deployed HCV RAG Gradio interface showing a query, generated answer, and retrieved FAISS evidence.

5. Conclusion

The reproduced implementation demonstrates a fully functional medical RAG system with high groundedness, low hallucination, and strong clinical correctness. All results were produced exclusively using our own code and evaluation pipeline, satisfying the Extra Credit requirements without duplicating any Baseline content.

References

S. Das, Y. Ge, Y. Guo, et al., "Two-Layer Retrieval-Augmented Generation Framework for Low-Resource Medical Question Answering Using Reddit Data: Proof-of-Concept Study," *Journal of Medical Internet Research*, vol. 27, p. e66220, 2025.

Z. Hammane, F. E. Ben-Bouazza, and A. Fennan, "Self-RewardRAG: enhancing medical reasoning with retrieval-augmented generation and self-evaluation in large language models," in *International Conference on Intelligent Systems and Computer Vision (ISCV)*, 2024.

Z. Ji, T. Yu, Y. Xu, N. Lee, E. Ishii, and P. Fung, "Towards mitigating LLM hallucination via self reflection," in *Findings of the Association for Computational Linguistics: EMNLP*, 2023, pp. 1827–1843.

S. Dhuliawala, M. Komelli, J. Xu, R. Raileanu, X. Li, and A. Celikyilmaz, "Chain-of-verification reduces hallucination in large language models," *arXiv preprint arXiv:2309.11495*, 2023.

G. Xiong, Q. Jin, Z. Lu, and A. Zhang, "Benchmarking retrieval-augmented generation for medicine," *arXiv preprint arXiv:2402.13178*, 2024.

J. Ge, S. Sun, J. Owens, V. Galvez, O. Gologorskaya, J. C. Lai, et al., "Development of a liver disease-specific large language model chat interface using retrieval augmented generation," *medRxiv*, 2023.

S. L. McNamara, P. H. Yi, and W. Lotter, "The clinician-AI interface: intended use and explainability in FDA-cleared AI devices for medical image interpretation," *NPJ Digital Medicine*, vol. 7, no. 1, p. 80, 2024.

S. Ghosh, U. Tyagi, S. Kumar, and D. Manoch, "BioAug: conditional generation based data augmentation for low-resource biomedical NER," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023, pp. 2533–2537.

S. Somani, S. Balla, A. W. Peng, R. Dudum, S. Jain, K. Nasir, et al., "Contemporary attitudes and beliefs on coronary artery calcium from social media using artificial intelligence," *NPJ Digital Medicine*, vol. 7, no. 1, p. 83, 2024.

M. Chaput, "Whoosh 2.7.4 documentation," 2024. [Online]. Available: https://whoosh.readthedocs.io/en/latest/

Teknium, theemozilla, karandd, and huemin_art, "Nous-Hermes-2-Mistral-7B-DPO," 2024. [Online]. Available: https://huggingface.co/NousResearch/Nous-Hermes-2-Mistral-7B-DPO