Retrieval-Augmented Generation for Medical Question Answering: A Comparative Study of Open-Source LLMs

Authors: Thomas Powell and Emiliano Vega

University of New Haven, CT, USA

DSCI-6004: Natural Language Processing

Abstract

This study presents a Retrieval-Augmented Generation (RAG) system aimed at improving medical question answering by combining open-source large language models (LLMs) with specialized medical information. We evaluate how well the LLaMA-3, Mistral-7B, and Phi-3 models perform on two different medical datasets: MedQuAD, which has expert-checked question and answer pairs, and MedlinePlus, a health information resource in XML format. Using a detailed evaluation method that includes ROUGE, BLEURT, and DeBERTa-based entailment scoring, along with expert reviews, we assess how well each model can produce answers that are correct, relevant, complete, and easy to understand. Results reveal that LLaMa-3 excels in our human evaluation, while Phi-3 performs better in our entailment scoring. The RAG framework demonstrably enhances the accuracy of the answers across all tested architectures.

1. Introduction

The increasing complexity and sensitivity of clinical question answering necessitate methodologies that combine robust language understanding with factual accuracy. Retrieval-Augmented Generation (RAG) provides a way to meet these needs by finding useful pieces of information and using them to create accurate responses. In this study, we set up a RAG system specifically for medical question answering, connecting it with well-known open-source large language models and testing how well they perform on different types of data. Our primary contributions are as follows: We have designed and deployed a flexible RAG system that interfaces with multiple open-source generative models, developed a comprehensive evaluation protocol that encompasses lexical similarity, semantic alignment, entailment fidelity, and expert quality judgments, and provided empirical insights into the interactions between model architecture, corpus structure, and downstream QA performance in medical domains.

2. RAG System Architecture

Our RAG system consists of five main components:

1. Corpus Processing: MedQuad is parsed from a CSV file into query-answer pairs and MedlinePlus content is extracted from an XML file. Both corpora are encoded by a sentence transformer, and an FAISS index is created from the embeddings.
2. Retrieval Engine: Queries are encoded and used to search for the

top-k documents from the corpus index.

3. Prompt Assembly: The LLMs use the retrieved documents as context, along with the query. Additional instructions are provided to increase generation quality.
4. Generation Layer: Prompts are passed on to one of the three LLMs using Hugging Face Transformers; there the LLM generates an answer to the given query using the provided context as well as pre-trained information.
5. Evaluation Pipeline: We evaluate generated outputs using automated metrics and manually rate them across predefined quality dimensions.

All modules are implemented in Python, leveraging data manipulation packages such as NumPy and Pandas; natural language processing packages like FAISS, transformers, sentence transformers, and PyTorch; and evaluation packages such as evaluate, ROUGE score, and BLEURT.

3. Domain-Specific Questions

We curated the following queries to encompass major clinical categories:

1. What is a stroke?
2. How does asthma affect the respiratory system?
3. What are the common causes of anemia?
4. How can someone reduce their risk of developing skin cancer?
5. What are the main differences between viral and bacterial infections?
6. What are the symptoms of depression?
7. What are the key signs of early-onset Alzheimer's disease?
8. How is Lyme disease transmitted?
9. What are the symptoms of HPV?
10. What should pregnant individuals know about prenatal vitamins?

These questions address acute, chronic, diagnostic, pharmacological, and preventive concerns relevant to general and specialized audiences.

4. Model Configurations

4.1. LLaMA-3

The Llama 3.2 collection of multilingual large language models (LLMs) provides ready-to-use generative models that adhere to instructions. They are available in sizes of 1 billion and 3 billion parameters (text in/text out). Optimized for multilingual conversation use cases, including agentic retrieval and summarizing activities, the Llama 3.2 instruction-tuned text-only models are above typical industry standards; they outperform many of the existing closed chat and open-source models. Using an enhanced transformer architecture, Llama 3.2 is an auto-regressive language model. Tuned versions match human preferences for usefulness and safety by use of supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF). We deployed the 3B parameter size and the instruction-tuned Llama 3.2 model locally using Hugging Face.

4.2. Mistral

The Mistral 7B model is based on a transformer architecture that utilizes sliding window attention (SWA) and grouped-query attention (GQA). GQA allows for greater batch sizes, thus more throughput, by considerably speeding the inference and lowering the memory need during decoding, which is crucial for real-time applications. SWA also aims to manage longer sequences more efficiently at a lower computing cost, addressing a typical LLM restriction. Together, these attention processes improve Mistral 7B's performance and efficiency. Mistral 7B makes a notable move toward balancing the objectives of maintaining big language model efficiency and upholding excellent performance. We deploy Mistral 7B using Hugging Face.

4.3 Phi-3

The Phi-3-Mini-128K-Instruct is a 3.8 billion-parameter, lightweight, state-of-the-art open model trained using the Phi-3 datasets. Emphasizing high-quality and reasoning-dense qualities, this dataset is comprised of filtered publicly accessible website data as well as synthetic data. The model is from the Phi-3 family; the Mini version comes in two variations, 4K and 128K, which indicate the context length (in tokens) it can support.

Following initial training, the model went through a post-training procedure, including supervised fine-tuning and direct preference optimization to improve its capacity to follow instructions and safety guidelines. When tested on common sense, understanding language, math, coding, remembering long-term information, and logical thinking, the Phi-3 Mini-128K-Instruct performed very well and was among the best models with fewer than 13 billion parameters. We deployed Phi-3-Mini-128k locally using Hugging Face.

5. Comparative Results of the Three LLMs

In total, there were 60 responses generated for our 10 queries: 30 per corpus, 20 per model, and 6 per question. Each of these responses was evaluated using various metrics. First, these responses received a Recall-Oriented Understudy for Gisting Evaluation (ROUGE) score. ROUGE compares the generated answers to a reference text and generates a score between 0 and 1 based on the similarity between the generated answer and the reference text. We used 4 different ROUGE scores: ROUGE-1, which evaluated the n-grams between the generated response and the reference text at n = 1; ROUGE-2, which looked at the answers when n = 2; ROUGE-L, which looks at the longest common sequence of words between the model's answer and the reference text; and ROUGE-L-SUM, which takes into account the whole text instead of just the longest common sequence.

Then the responses were evaluated using BLEURT. BLEURT takes input lists of predicted sentences and reference sentences and outputs a score between 0 and approximately 1 for each individual sentence prediction. BLEURT is used to indicate how similar the generated text is to the reference text and is reported to be better suited to human judgment compared to BERT.

Next, the responses underwent a Natural Language Inference (NLI) evaluation using DeBERTa (Decoding-enhanced BERT with

Disentangled Attention) MNLI. We used DeBERTa to evaluate the faithfulness and confidence of the responses. Faithfulness shows how much the answer matches the original text, where entailment means the answer is backed by the text, contrast means the answer contradicts the text (which is a mistake), and neutral means it's not clear if the answer is supported by the text. Confidence scores indicate how confident the model is that the answer it generated is correct.

Finally, the model-generated responses were sent to Ally Tran, PA-SII, a physician's assistant undergoing clinicals at Yale, to be evaluated by a professional in the medical field. She evaluated the responses based on 4 criteria: factual accuracy, relevance, completeness, and clarity. Factual accuracy evaluates whether the model response is medically correct. Relevance evaluates whether the model stays on topic. Completeness evaluates whether the model answers the questions fully. Clarity evaluates whether the model-generated response is easy to read and understand. The responses were evaluated on a scale of 0 to 3. 0 would indicate that the model completely failed or only minimally met the criteria. 1 would indicate that the model only partially met the criteria. 2 would indicate that the model mostly met the criteria. 3 would indicate that the response perfectly met the criteria. All these evaluations were averaged and plotted to compare the average evaluation for each question, model, and corpus.

5.1 Evaluations by Question

According to the ROUGE scores, the strongest results on average were for Questions 3, 8, and 10, with Questions 4, 5, and 7 generating the lowest scoring answers on average. (Figures 1–4) The BLEURT evaluation gave Question 1 the only positive score, while all other answers received negative scores. Questions 4 and 9 had the lowest scores among them. (Figure 5) The DeBERTa evaluations gave a very positive result, with none of the questions receiving a response that was evaluated to be a contrast from the reference text. Questions 3 and 10 received the entailment label for all 6 of the answers generated. Questions 1, 2, 4, and 8 also performed very strongly. Question 7 leaned in a more neutral direction, indicating that it was unclear whether most of the responses were aligned with the reference text. (Figure 6) With regards to factual accuracy, Tran scored Questions 1, 2, and 8 very highly, as all 6 generated responses for Question 2 received a factual accuracy score of 3. (Figure 7) Question 4 got an equal amount of 0s and 3s and was the only question to receive multiple 0s. She determined Questions 1, 2, and 3 to have received the most relevant responses, with the responses for Question 4 again receiving the lowest evaluation. (Figure 8) Questions 1, 2, and 3 had the most complete responses, with Questions 4, 5, and 10 receiving the lowest scores. (Figure 9) The responses for Questions 1, 2, and 8 were rated the highest by the PA student evaluator for clarity, with Questions 4, 5, 6, and 10 receiving the lowest rated responses. (Figure 10)

5.2 Evaluations by Model

On average, Phi-3 received the highest ROUGE scores for its generated responses, with LLaMa-3 slightly outperforming

Mistral-7B. (Figures 11-14) All three models had a negative BLEURT score, with Phi-3 being the closest to zero. LLaMa-3 again performed better than Mistral-7B. (Figure 15) Phi-3 also received the highest evaluation from DeBERTa, with 80% of its responses receiving the entailment label for faithfulness. Here, Mistral-7B outperformed LLaMa-3, receiving 65% entailment compared to 50% for LLaMa-3. (Figure 16) The human evaluations, however, give a different picture. The evaluator gave Mistral-7B and LLaMa-3 equal factual accuracy scores on average, both rated higher than Phi-3. (Figure 17) She determined that Mistral-7B had the most relevant responses on average, followed by LLaMa-3 and Phi-3. (Figure 18) LLaMa-3 had the most complete responses, followed by Phi-3, with Mistral-7B ranked at the bottom. (Figure 19) LLaMa-3 also received the highest evaluation for clarity, followed by Mistral-7B and Phi-3. (Figure 20)

5.3 Evaluation by Corpus

Across all the ROUGE metrics, the responses using the MedQuAD corpus performed better than the responses using the MedlinePlus corpus. (Figures 21-24) Both BLEURT scores were negative, with MedlinePlus outperforming MedQuAD on this metric. (Figure 25) MedQuAD-based responses received more entailment labels from DeBERTa than MedlinePlus, which was more of an even split. (Figure 26) The PA student evaluator found that the answers from the MedlinePlus collection were better than those from the MedQuAD collection in all four areas: factual accuracy, relevance, completeness, and clarity. (Figures 27-30)

6. Analysis

The evaluation reveals nuanced trade-offs across the three models. On the MedQuAD dataset, Phi-3 achieved significantly higher ROUGE-1 (0.562) and ROUGE-2 (0.523) scores than both LLaMA-3 and Mistral-7B. This result reinforces its strength in generating text that aligns closely—at the lexical level—with structured, templated reference answers. However, this strong similarity at the surface level did not lead to better meaning or factual accuracy: Phi-3 had the lowest score in BLEURT (-0.569) and the lowest factual accuracy (1.8/5), showing it depended more on simple pattern matching instead of thoughtful reasoning.

LLaMA-3 demonstrated stronger BLEURT scores (-0.568) on MedlinePlus and relatively better factual accuracy in both datasets, scoring 2.1 in MedQuAD and 2.4 in MedlinePlus. It also achieved the highest clarity rating (2.4) on MedlinePlus. Mistral-7B, despite a BLEURT score of -0.837 on MedQuAD and -0.591 on MedlinePlus, had the highest overall factual accuracy rating on MedlinePlus (2.5) and was consistently rated more relevant (2.9) and complete (1.7) than its counterparts in semi-structured settings.

The findings suggest that while ROUGE remains sensitive to surface-level overlap, it poorly captures deeper qualitative attributes of generation. BLEURT and human scores, by contrast, better reflect fluency, informativeness, and grounding—crucial traits in the medical QA domain.

7. Discussion

Each model offers distinct operational strengths:

Phi-3 demonstrates compelling performance on lexically predictable, template-based queries but underperforms on factual accuracy and semantic grounding. Its utility is best reserved for constrained QA environments.

LLaMA-3 presents a balanced profile across metrics, excelling in clarity and BLEURT on MedlinePlus. This makes it a strong candidate for tasks requiring higher semantic precision, though its computational requirements are considerable.

Mistral-7B emerged as the most grounded model in human assessments, particularly on factuality and relevance. Its relatively strong performance under semi-structured retrieval suggests robustness to noise and flexibility across corpora.

These findings affirm that generation quality is strongly mediated by corpus structure and retrieval conditioning. Notably, the gap between lexical and semantic alignment metrics underscores the importance of diversified evaluation for domain-sensitive tasks like medical QA.

8. Conclusion

This study provides a rigorous comparison of open-source LLMs integrated into a RAG pipeline for medical QA. We demonstrate that retrieval structure, model architecture, and query formulation collectively impact generation quality. While Phi-3 performs admirably under templated queries, LLaMA-3 and Mistral-7B exhibit greater semantic resilience and factual grounding.

Future work will explore cross-lingual retrieval, clinical note integration, and dynamic prompt optimization informed by real-time user feedback.

9. Figures

9.1. Question Evaluation

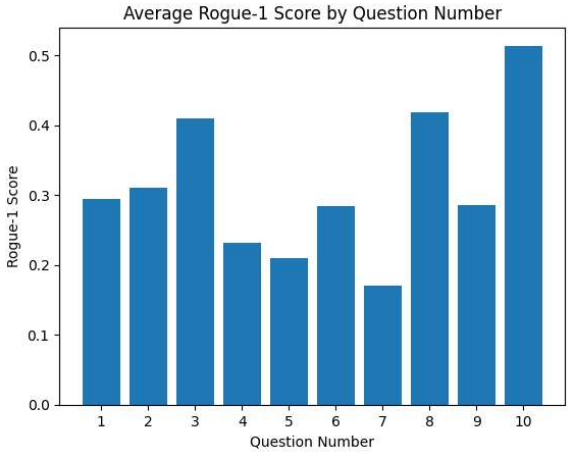

Figure 1: A bar chart consisting of average Rogue-1 scores for each question.

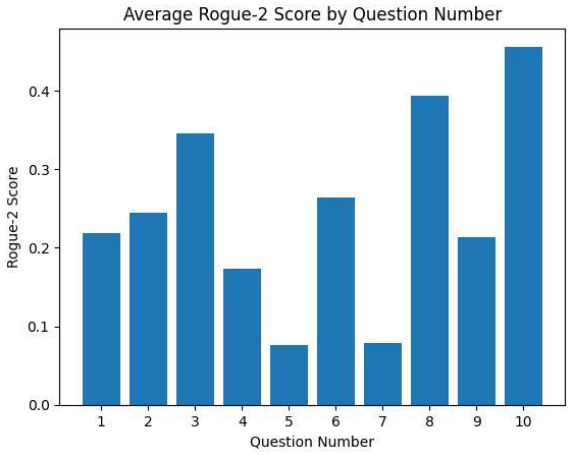

Figure 1: A bar chart consisting of average Rogue-2 scores for each question.

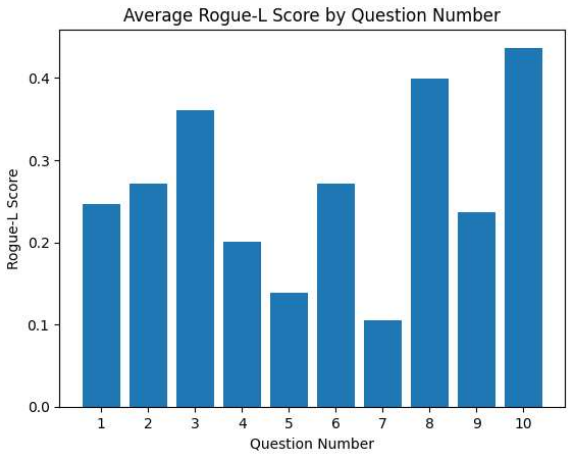

Figure 3: A bar chart consisting of average Rogue-L scores for each question.

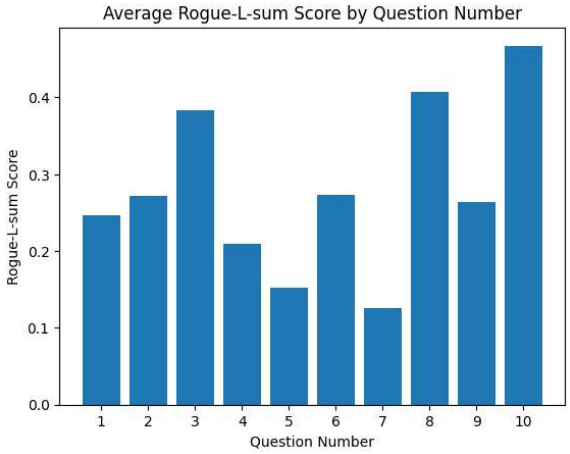

Figure 4: A bar chart consisting of average Rogue-L-sum scores for each question.

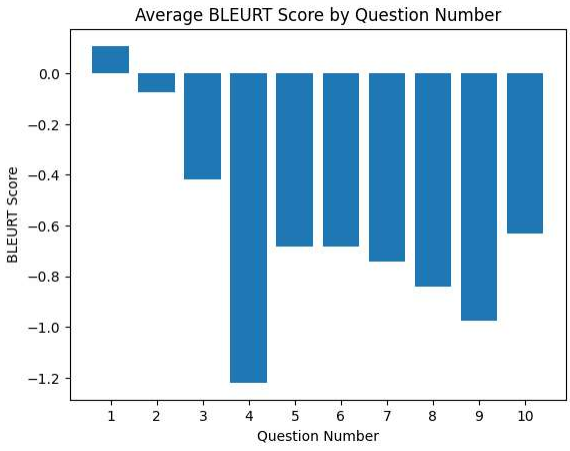

Figure 5: A bar chart consisting of average BLEURT scores for each question.

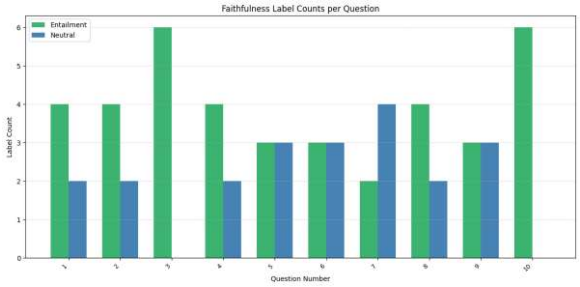

Figure 6: A clustered bar chart consisting of DeBERTA faithfulness label counts for each question.

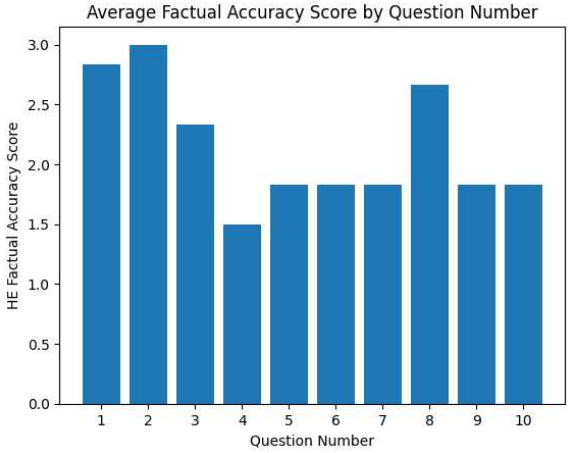

Figure 7: A bar chart consisting of average expert-evaluated factual accuracy scores for each question.

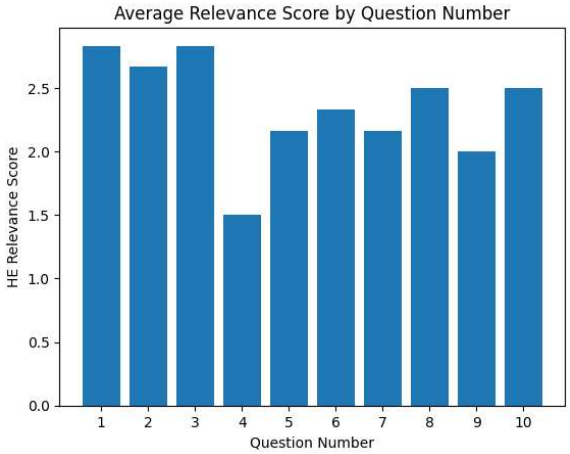

Figure 8: A bar chart consisting of average expert-evaluated relevancy scores for each question.

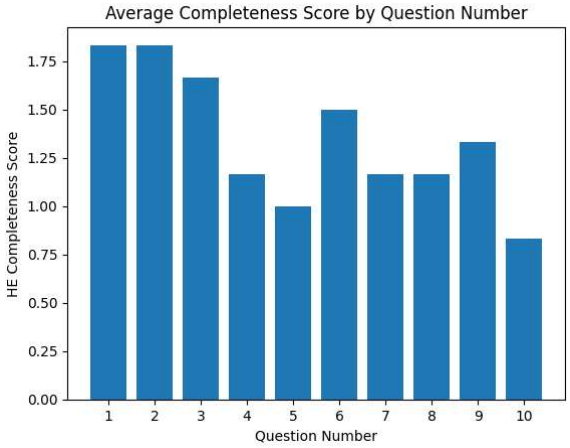

Figure 9: A bar chart consisting of average expert-evaluated completeness scores for each question.

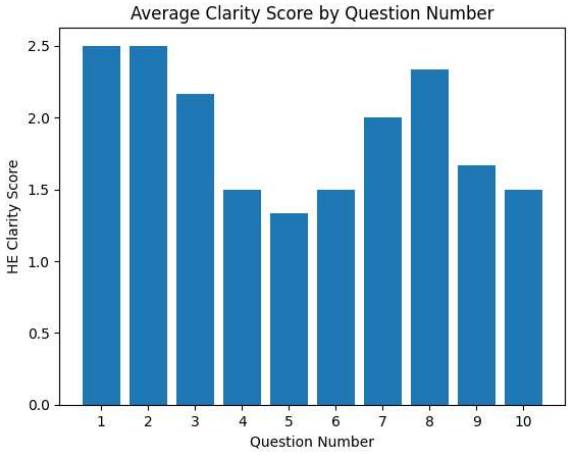

Figure 10: A bar chart consisting of average expert-evaluated clarity scores for each question.

9.2. Model Evaluation

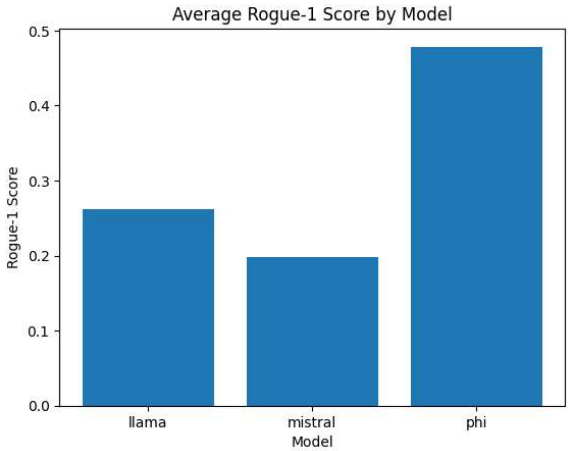

Figure 11: A bar chart consisting of average Rogue-1 scores for each model.

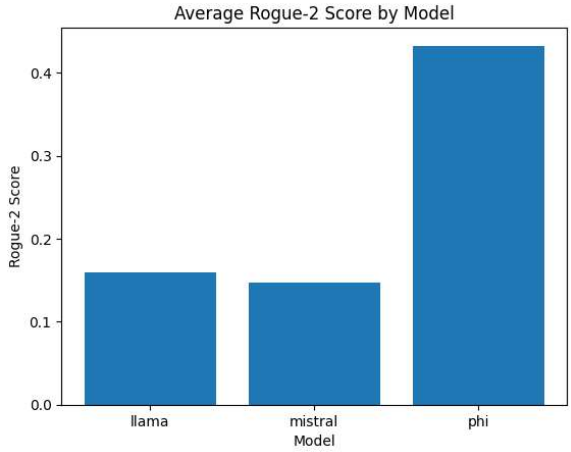

Figure 12: A bar chart consisting of average Rogue-2 scores for each model.

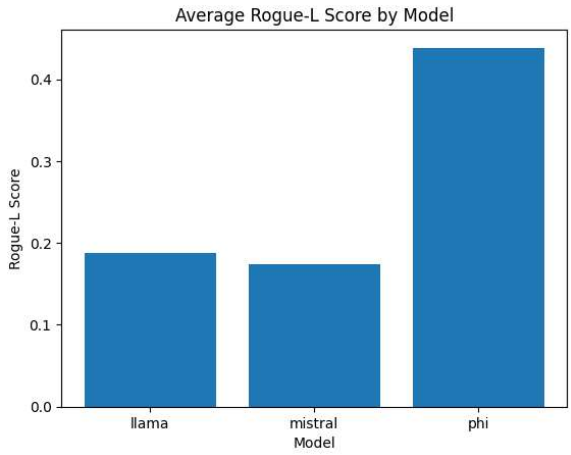

Figure 13: A bar chart consisting of average Rogue-L scores for each model.

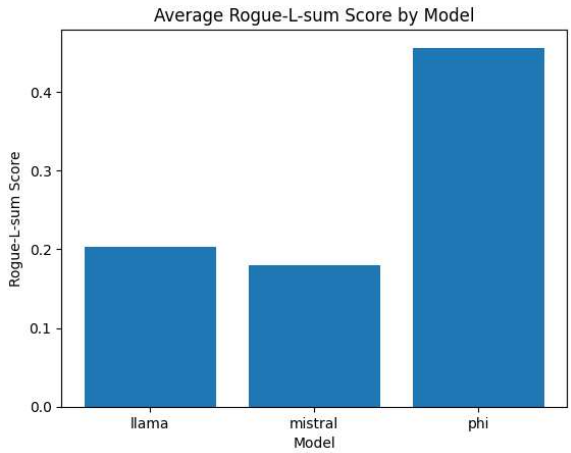

Figure 14: A bar chart consisting of average Rogue-1 scores for each model.

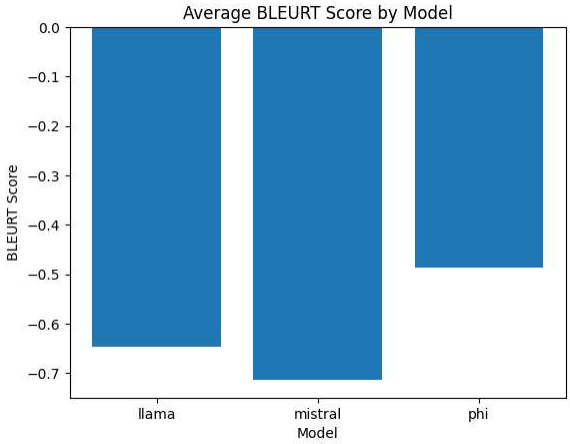

Figure 15: A bar chart consisting of average BLEURT scores for each model.

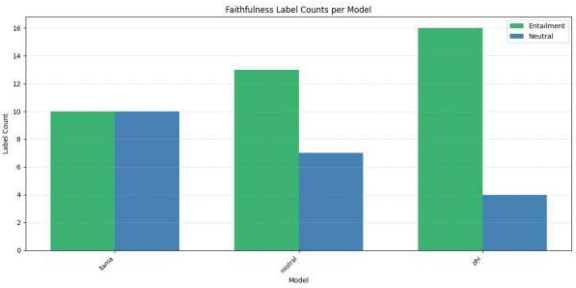

Figure 16: A clustered bar chart consisting of DeBERTA faithfulness label counts for each model.

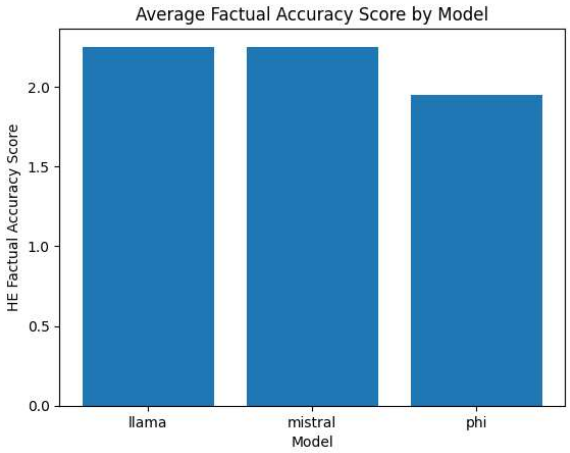

Figure 17: A bar chart consisting of average expert-evaluated factual accuracy scores for each model.

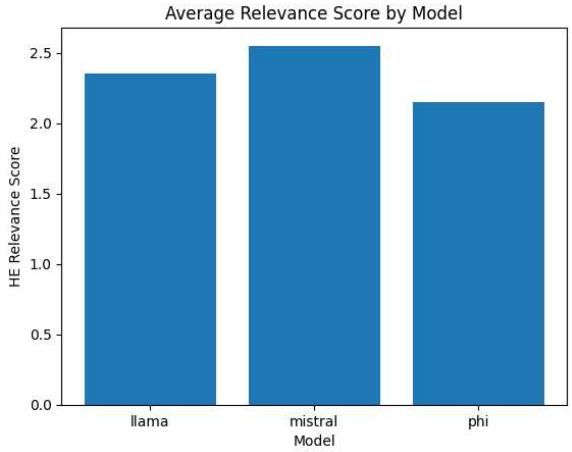

Figure 18: A bar chart consisting of average expert-evaluated relevance scores for each model.

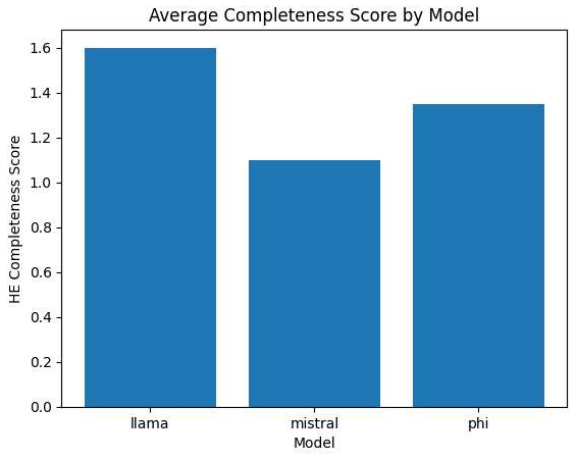

Figure 19: A bar chart consisting of average expert-evaluated completeness scores for each model.

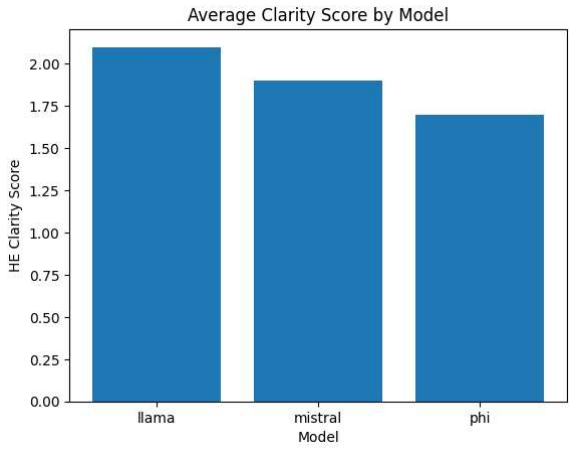

Figure 20: A bar chart consisting of average expert-evaluated clarity scores for each model.

9.2. Corpus Evaluation

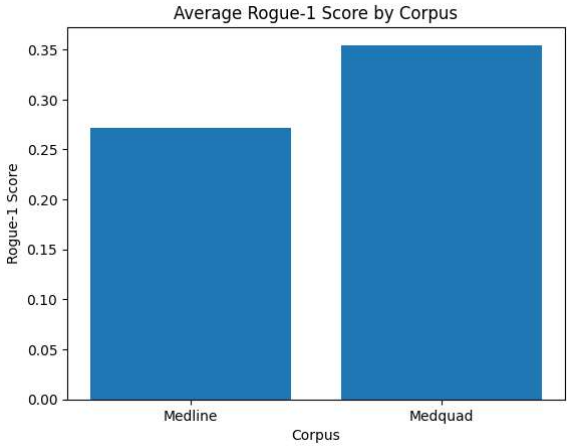

Figure 21: A bar chart consisting of average Rogue-1 scores for each corpus.

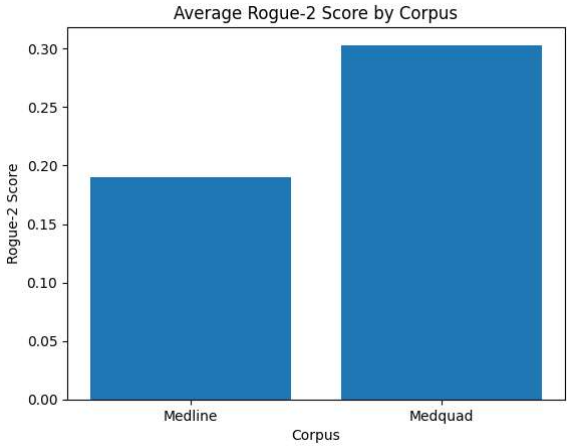

Figure 22: A bar chart consisting of average Rogue-2 scores for each corpus.

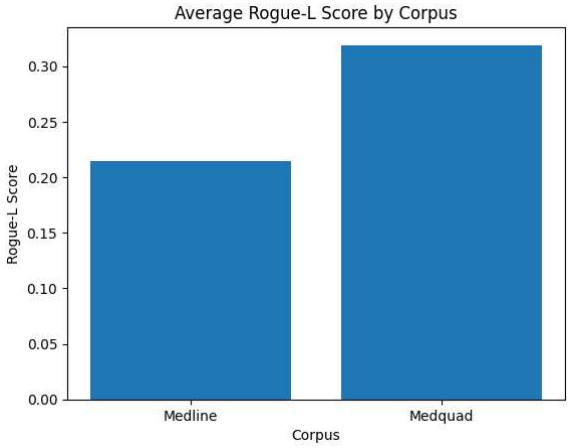

Figure 23: A bar chart consisting of average Rogue-L scores for each corpus.

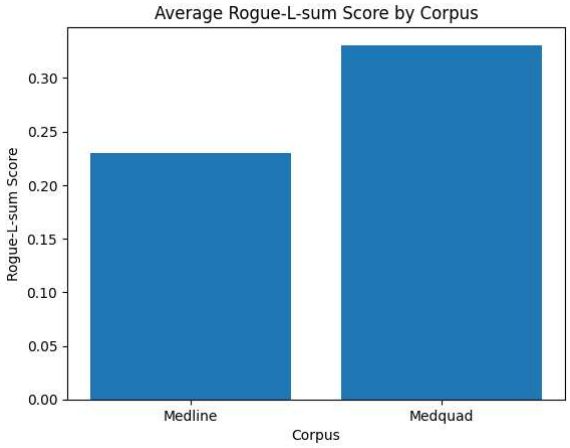

Figure 24: A bar chart consisting of average Rogue-L-sum scores for each corpus.

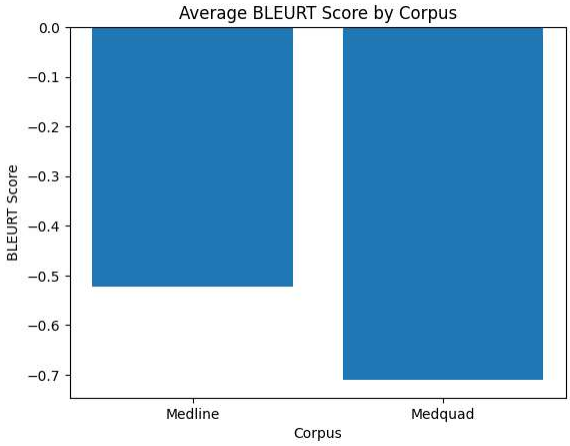

Figure 25: A bar chart consisting of average BLEURT scores for each corpus.

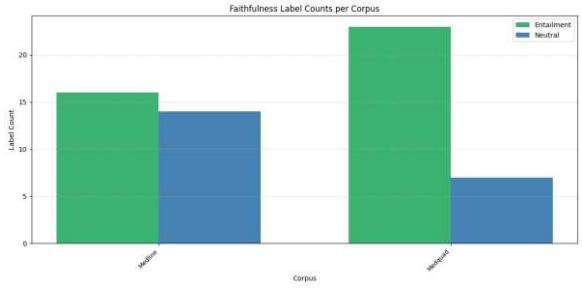

Figure 26: A clustered bar chart consisting of DeBERTA faithfulness label counts for each corpus.

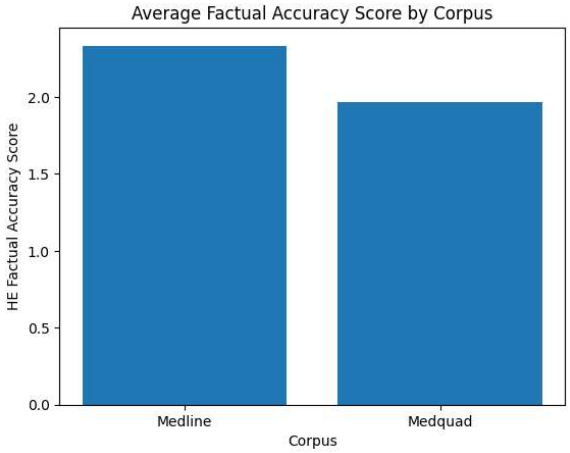

Figure 27: A bar chart consisting of average expert-evaluated factual accuracy scores for each corpus.

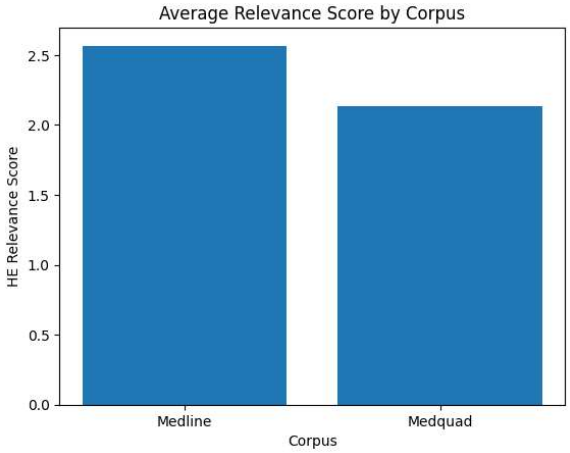

Figure 28: A bar chart consisting of average expert-evaluated relevance scores for each corpus.

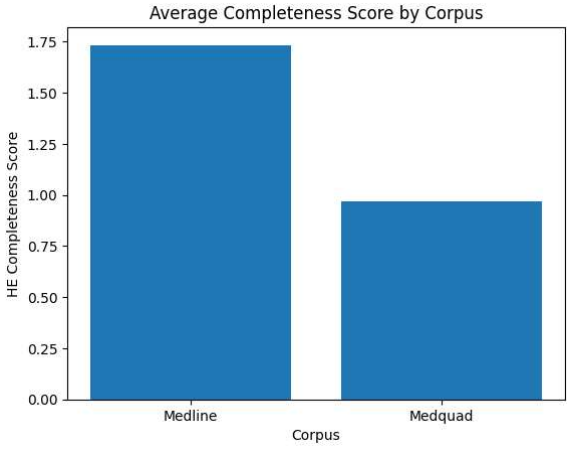

Figure 29: A bar chart consisting of average expert-evaluated completeness scores for each corpus.

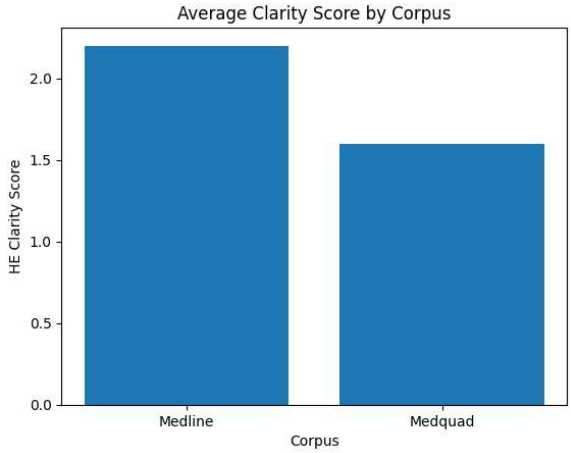

Figure 30: A bar chart consisting of average expert-evaluated clarity scores for each corpus.

References

Ben Abacha, A. and Demner-Fushman, D. (2019) 'A question-entailment approach to question answering', BMC Bioinformatics, 20(1). doi:10.1186/s12859-019-3119-4.

Bilenko, M. (2024) 'New models added to the Phi-3 family, available on Microsoft Azure', Microsoft, 21 May. Available at: https://azure.microsoft.com/en-us/blog/new-models-added-to-the-phi-3-family-available-on-microsoft-azure/ (Accessed: 02 May 2025).

Devlin, J. et al. (2019) 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1, pp. 4171–4186. doi:10.18653/v1/n19-1423.

Douze, M. et al. (2025) The faiss library, arXiv.org. Available at: https://doi.org/10.48550/arXiv.2401.08281 (Accessed: 02 May 2025).

He, P. et al. (2021) DeBERTa: Decoding-enhanced Bert with disentangled attention, arXiv.org. Available at: https://doi.org/10.48550/arXiv.2006.03654 (Accessed: 02 May 2025).

Jiang, A.Q. et al. (2023) Mistral 7B, arXiv.org. Available at: https://doi.org/10.48550/arXiv.2310.06825 (Accessed: 02 May 2025).

Lee, J. et al. (2019) 'BioBERT: A pre-trained biomedical language representation model for biomedical text mining', Bioinformatics, 36(4), pp. 1234–1240. doi:10.1093/bioinformatics/btz682.

Lin, C.-Y. (2004) 'ROUGE: A Package for Automatic Evaluation of Summaries', Text Summarization Branches Out, pp. 74–81. doi:https://aclanthology.org/W04-1013/.

Singhal, K. et al. (2025) 'Toward expert-level medical question answering with large language models', Nature Medicine, 31(3), pp. 943–950. doi:10.1038/s41591-024-03423-7.

Touvron, H. et al. (2023) Llama: Open and efficient foundation language models, arXiv.org. Available at: https://doi.org/10.48550/arXiv.2302.13971 (Accessed: 02 May 2025).

XML files (2025) MedlinePlus. Available at: https://medlineplus.gov/xml.html (Accessed: 02 May 2025).