BioRAGLLM: Evaluating Large Language Models with Retrieval-Augmented Generation

Andrew Cai andcai@mit.edu
Gustavo Ramirez ramgus@mit.edu
Eddy Ogola ogolaedd@mit.edu
Tinah Hong tinahh@mit.edu

Abstract

Large language models are becoming increasingly popular. An obvious practical use case, is long-winded interactions. However, this is a difficult problem since KV caching previous tokens is limited by memory, therefore limiting quality LLM output to the pretraining length. Prior work on length extrapolation has been done, but they're still not ideal. Particularly, StreamingLLM [1] is perfomant with low perplexity but evicts some tokens which might be important for the task at hand. Retrieval Augmented Generation [2] is a potential solution to this problem as it indexes specific documents helpful for a given task. This project, first observes the weakness of StreamingLLM then explores the performance of RAG across four different small, open-source models run locally on an edge device. We curate a dataset for Biology Q&A based on content from MIT's undergraduate General Institute Requirement(GIR) and build an LLM system for this task which we call BioRAGLLM. We observe that Mistral is the best model, with RAG, for this task, and also point out RAG's weaknesses in solving the problem of indexing long text datasets.

1 Introduction

The emergence of Large Language Models (LLMs) in Natural Language Processing (NLP) has marked a paradigm shift, enabling sophisticated text generation and complex language understanding. Despite their advancements, LLMs face a significant challenge in sustaining quality output during prolonged interactions, primarily due to memory limitations in Key-Value (KV) caching of tokens.

This paper explores the application of Retrieval-Augmented Generation (RAG) in overcoming these constraints. RAG enhances LLMs by integrating document retrieval into the generation process, a method particularly useful for tasks involving extensive text references, such as Biology Question and Answering (QA).

Our study introduces BioRAGLLM, a system employing RAG to boost the capabilities of four distinct small, open-source models in a local edge computing environment. We specifically focus on Biology QA, utilizing a dataset curated from MIT's undergraduate General Institute Requirement (GIR) content, to test the efficacy of this approach.

We begin with a review of relevant literature in LLMs and RAG, followed by an in-depth description of our methodology, encompassing dataset preparation, model selection, and RAG implementation. Subsequent sections present our findings, analyze the performance of different models with RAG, and discuss its strengths and limitations in handling extensive text datasets. Finally, the paper concludes by proposing potential future research directions in this domain.

2 Related Work

The field of AI language models has seen significant advancements in efforts to extend context windows, enhancing the coherence and engagement of conversations. A prominent development

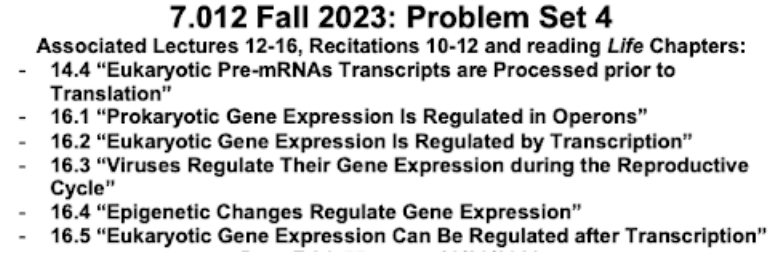

Figure 1: Pset 4 chapter readings

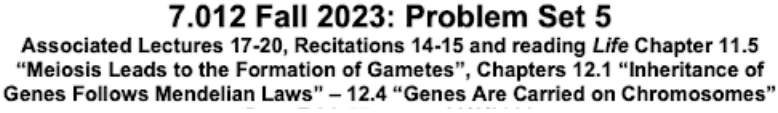

Figure 2: Pset 5 chapter readings

is the Streaming Large Language Model (StreamingLLM) by Xiao et al. (2023), which efficiently handles extended context in real-time data streams. Despite its effectiveness, StreamingLLM faces a limitation in its inability to attend to tokens once they are evicted from the memory, leading to potential gaps in context retention.

To address this limitation, the integration of StreamingLLM with Retrieval-Augmented Generation (RAG) frameworks like LlamaIndex presents a promising solution. This combination allows the model to fetch evicted information back into the current context, thereby producing more coherent text while maintaining access to past interactions and external databases.

In addition, the work of Jiang et al. (2023) on Mistral 7B introduces the use of sliding window attention mechanisms in smaller models. This approach is particularly effective for processing long sequences without extensive memory requirements, a key factor in the performance of models in constrained environments. Mistral 7B demonstrates how sliding window attention can be leveraged to enhance the capabilities of LLMs, especially when dealing with lengthy texts.

The studies by Jiawei Chen et al. (2023) and Victoria Lin et al. (2023) further contribute to this field by evaluating the performance of RAG within LLMs and proposing methods for fine-tuning both the LLM and retrieval components. These advancements collectively signify the potential of RAG in conjunction with streaming LLMs, paving the way for more advanced and user-centric conversational AI systems.

3 Methodology

Here we outline how we went about curating the dataset for BioRAGLLM, the models we used for this task, and the tools we used for the RAG pipeline.

3.1 Dataset

Introduction to Biology, popularly called 7.012, is a General Institute Requirement(GIR) for undergraduates. A bulk of the work for the class is problem sets, which have a particular structure that made them ripe for our project curiosity. The class' assigned textbook is Life: The Science of Biology[8]. Each problem set comes along with a set of chapters, meant to be the reference material for the problems. The problems are also set up conveniently, with most of them being multi-choice, with the exception of only a few asking for a short answer.

We cherry-picked 3 problem sets that had the fewest diagrams: PSets 4, 5, and 8 from Fall 2023. The assigned chapters were as shown in the figures below.

7.012 Fall 2023: Problem Set 8

Associated Lectures 28-30, Recitations 22-23 and reading *Life* Chapter 19.1 "The Four Major Processes of Development Are Determination, Differentiation, Morphogenesis, and Growth" as well as the following chapters from Chapter 41

- **41.1 "Animals Use Innate and Adaptive Mechanisms for Defense"**
- **41.2 "Innate Defenses Are Nonspecific"**
- **41.3 "Adaptive Defenses Are Specific"**
- **41.4 "The Humoral Adaptive Response Involves Antibodies"**
- **41.5 "The Cellular Adaptive Response Involves T Cells and Receptors"**

Figure 3: Pset 8 chapter readings

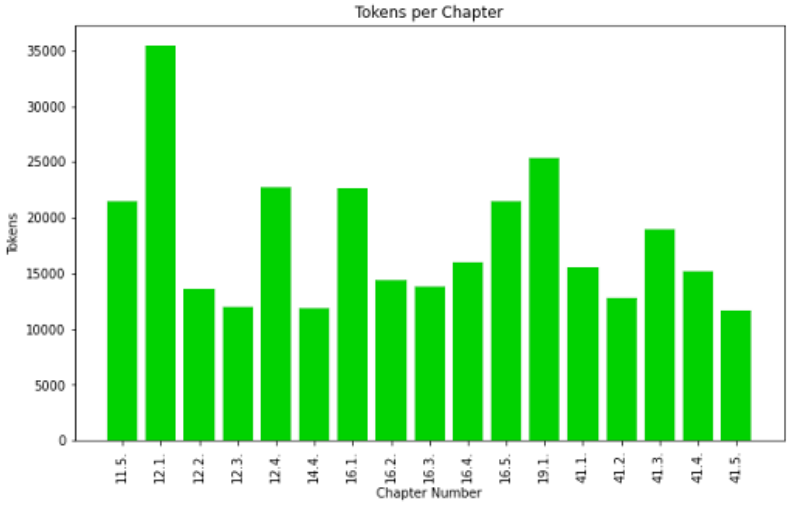

Figure 4: Enter Caption

As we can see from the chart below, token counts for the different chapters was pretty high. With a low of 10k tokens to a maximum of 35k tokens. Altogether, some psets required referencing over 30k, to a high of 100k tokens.

Dataset curation We only picked the multi-choice questions and obtained the ground truth answers from the published answer keys. To make the choices and answers more natural, we tweaked the choice number from just being, for example 1) prokaryote 2) eukaryote, to choice 1 of 2) prokaryote, choice 2 of 2) eukaryote. The questions were left as is, only to be prepended with the prompt engineering line: "Give me the answer to the following question given the following choices."

We manually extract text from the required chapters for each problem set, of course, ignoring all graphics.

3.2 Models

In spirit with the class' goal of running big deep learning models on edge devices, we ran large, but relatively small, open-sourced large language models on an Apple Macbook Pro M2 Max device. The specific models were sourced from Ollama.ai[9].

LLama 2 7B [4] chat Fine-tuned for dialogue/chat.

Mistral 7.3B [6] Uses Sliding Window Attention for long sequences.

Vicuna 7B [5] Fine-tuned for chat.

Orca 2 7B [3] Fine-tuned for chat; Designed to excel particularly well in reasoning.

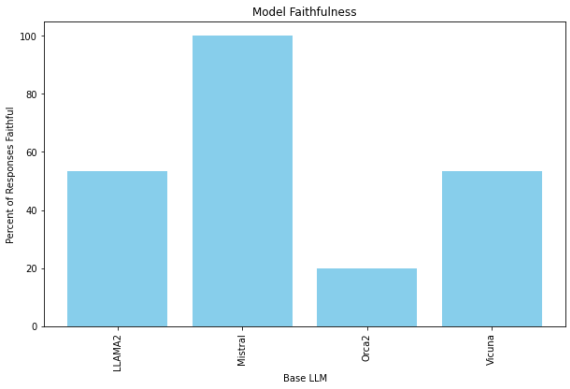

Figure 5: Automatic evaluation of faithfulness of each base LLM's responses.

3.3 Retrieval Augmented Generation

The RAG pipeline is standard with the documents stored in a directory database. The database is then split into tokens, chunked, and then embedded using the 'BAAI/bge-small-en-v1.5' HuggingFace embedding model. This embedding model, in addition to the large language model in question, are used to construct and indexed database('KeywordTableIndex' speficically, in LlamaIndex). The index is then used to build a query engine, which takes in queries, searches for relevant embedding points in the index, and then uses the LLM's context to respond.

Task setup We loaded the models from Ollama, and used Llama Index for the RAG pipeline, chunking, embedding, and indexing the text in the database(reference chapters), for each problem set.

4 Results

As described in our methodology, we evaluate each of our four base LLM models on our 7.012 problem set dataset. We consider three metrics: Faithfulness, relevancy, and correctness. The former two are evaluated using LlamaIndex's automated `FaithfulnessEvaluator` and `RelevancyEvaluator` modules. Each response is evaluated with a boolean value (0 or 1) against each metric, hence the model's overall performance on the metric is computed as the percentage of responses that satisfy the metric. Meanwhile, we evaluate correctness by manually grading each response against the answer key, using the same metrics as a real-world 7.012 grader. Faithfulness and relevancy are used as auxiliary metrics to gauge the performance of the RAG referencing and LLM integration, respectively, of each model. However, correctness is the sole metric that evaluates the true performance of our models on the Biology Q&A task.

Faithfulness Faithfulness evaluates if a response is faithful to its contexts. We observe each model's faithfulness in Figure 5, where Mistral far exceeds the other LLMs, with each of its responses deemed as faithful. Individual responses from Mistral accurately demonstrate this, citing specific instances of the textbook for many answers.

Relevancy Relevancy evaluates if a response is relevant to the question or task at hand. Again, as observed in Figure 6, Mistral performs above other LLMs at 80%, though LLAMA2 and Vicuna perform relatively well.

Correctness Correctness evaluates the true performance of the model on the Biology Q&A task. Ultimately, it evaluates which base LLM performs the best, and the efficacy of our BioRAGLLM model. As observed in Figure 7, Mistral answers 66.7% of test set questions correctly, while the other models struggle to answer any questions.

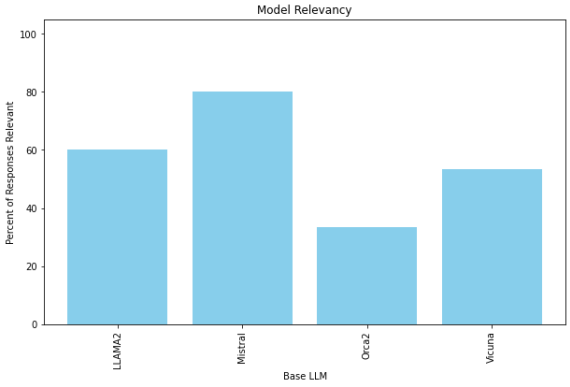

Figure 6: Automatic evaluation of relevancy of each base LLM's responses.

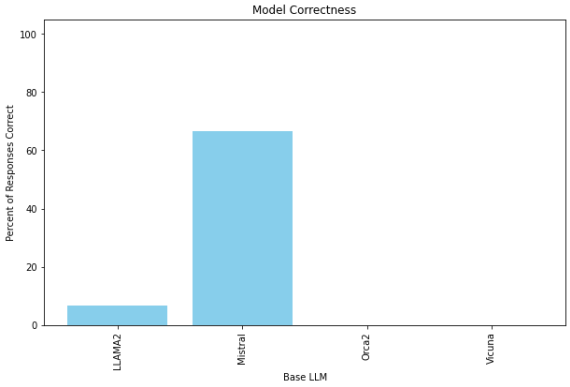

Figure 7: Manual grading of each base LLM's responses on provided answer key.

5 Discussion

The comparative analysis of models in the Biology Q&A task reveals a clear superiority of Mistral over its counterparts. This is evident not only in the raw performance metrics but also in the qualitative aspects of the responses. Other models, despite their ability to generate contextually relevant content, fail to maintain a high level of faithfulness to the provided context. This results in a substantial discrepancy in factual accuracy, as evidenced by the faithfulness metric, where these models only align with the context about half of the time.

A critical observation from our analysis is the limitation of the automated LlamaIndex relevancy metric as a proxy for response relevance. Most models tend to generate content that, while topically relevant, often fails to address the specific query posed. This is a common issue where the responses, albeit related to the broader biological subject, are tangentially connected or completely unrelated to the actual question. Furthermore, even when models do provide relevant information, they often embed the actual answer within a verbose response rather than presenting it concisely and directly.

Mistral's dominance in this task is attributed to its ability to generate precise and direct answers. Unlike other models that veer into tangentially related content or bury the answer within lengthy passages, Mistral consistently provides clear and specific responses to each question.

Looking ahead, a promising direction for our research is the integration of our RAG pipeline with StreamingLLM. The ultimate goal of this integration is to leverage the advancements made in attention sinks by StreamingLLM while addressing the challenge of evicted tokens. By combining StreamingLLM's efficient handling of extended contexts with our RAG pipeline's capability to fetch relevant information, we can potentially create a system that not only generates infinitely coherent text but also retains continuous access to past information. This synergy could overcome the current

limitations observed in StreamingLLM and other models, leading to a more robust and reliable system for complex tasks like Biology Q&A.

In essence, the fusion of StreamingLLM's architecture with our enhanced RAG framework promises to be a significant leap forward in the field. It could pave the way for models that not only understand and generate natural language with high precision but also maintain a seamless continuity of context over extended interactions.

6 Conclusion

This study has demonstrated that Retrieval-Augmented Generation (RAG) serves as a potent augmentation for autoregressive large language models (LLMs), especially in tasks demanding references to extensive text datasets, such as question and answering. Our empirical investigation highlights the differential impact of RAG across various models, with Mistral 7B emerging as a significantly superior performer compared to its counterparts, namely Orca 2 7B, Llama 2-chat 7B, and Vicuna 7B. Mistral 7B's proficiency in directly addressing queries, coupled with its integration with RAG, has set a new benchmark in the domain of LLMs.

However, the application of RAG is not without its challenges. One of the primary limitations observed is the time-intensive nature of the indexing phase, which scales with the length of the database. This aspect poses a bottleneck in terms of efficiency, especially when handling large-scale datasets. Additionally, the current implementation of RAG, when applied without fine-tuning, can occasionally lead models astray, suggesting that while RAG provides a directional benefit to LLMs, it requires careful calibration to achieve optimal performance.

The insights gained from this study point towards the necessity for finer adjustments and more nuanced applications of RAG, particularly when integrating it with different LLM architectures. Future research should focus on optimizing the indexing process to enhance efficiency and exploring ways to fine-tune RAG-augmented models for specific tasks. The goal is to harness the full potential of RAG in enhancing the accuracy and relevance of LLM outputs, especially in complex and information-dense fields such as biology QA.

In conclusion, the integration of RAG with LLMs represents a significant step forward in the field of natural language processing. The advancements made through this research not only improve the capability of LLMs in handling extended text references but also open new avenues for further enhancements in model performance and applicability.

References

[1] Xiao, G., Tian, Y., Chen, B., Han, S., & Lewis, M. (2023). Efficient streaming language models with attention sinks. arXiv preprint arXiv:2309.17453.

[2] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

[3] Mitra, A., Del Corro, L., Mahajan, S., Codas, A., Simoes, C., Agarwal, S., ... & Awadallah, A. (2023). Orca 2: Teaching Small Language Models How to Reason. arXiv preprint arXiv:2311.11045.

[4] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

[5] Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., & Xing, E. P. (2023, March). Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90

[6] Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. D. L., ... & Sayed, W. E. (2023). Mistral 7B. arXiv preprint arXiv:2310.06825.

[7] MIT. (2004). 7.012 Introduction to Biology (Fall 2004). Massachusetts Institute of Technology. https://ocw.mit.edu/courses/7-012-introduction-to-biology-fall-2004/

[8] Sadava, D., Hillis, D. M., Heller, H. C., & Berenbaum, M. (2019). Life: The Science of Biology (11th ed.). Sinauer Associates, Inc; W.H. Freeman and Company.

[9] jmorganca. (2023). Ollama (Version 0.1.15) [Software]. Available from https://github.com/jmorganca/ollama

[10] Liu, J. (2022). LlamaIndex [Software]. $https://doi.org/10.5281/zenodo.1234$. Available at $https://github.com/jerryjliu/llama\_index$

[11] J. Chen et al. (2023). Benchmarking Large Language Models in Retrieval-Augmented Generation. Advances in Neural Information Processing Systems. https://arxiv.org/abs/2309.01431

[12] X. V. Lin et al. (2023). RA-DIT: Retrieval-Augmented Dual Instruction Tuning. arXiv preprint arXiv:2310.01352. https://arxiv.org/abs/2310.01352