RAG-IR Augmented LLMs

Irven Minhas and **Alan Zhao** and **Junze Wei** and **Richie Hsieh** and **Yoshio Kondoh**
University of Toronto Mississauga

Abstract

This document investigates the effectiveness of combining Retrieval-Augmented Generation (RAG) and Information Retrieval (IR) in reducing hallucinations in large language models (LLMs). Although great at most tasks, when lacking an answer, they often produce incorrect or misleading information. By integrating IR with RAG, we aim to improve accuracy, reduce illusion rates, and enhance consistency in question-answering tasks. We hypothesized that grounding LLM outputs in verifiable external sources would significantly enhance their reliability and trustworthiness. This document is a showcase of our of our methodology, results, and improvements.

1 Introduction

1.1 Research Question

How effective are Retrieval-Augmented Generation (RAG) and Information Retrieval (IR) in reducing hallucinations in large language models (LLMs) such as Deepseek, Gemma-3, and Llama 3, particularly in terms of factual accuracy, and source consistency in question-answering tasks?

1.2 Background Research

Illusions in LLMs refer to instances where the model generates plausible but factually incorrect or unsupported information. RAG integrates external knowledge retrieval with generative models, potentially reducing hallucinations by grounding responses in verified external sources.

IR systems, which retrieve relevant documents from large databases, can be integrated with RAG to ground model response in information pulled from verified sources. This project will compare the performance of LLMs with and without IR-enhanced RAG to assess its effectiveness in improving factual accuracy and reducing hallucinations.

1.3 Hypothesis

We expect IR-RAG models to achieve higher retrieval precision, accuracy, and response reliability, particularly in domains requiring precise factual information.

We will compare the performance of LLMs with and without IR-enhanced RAG on a set of QA tasks. The baseline will be the standard LLM performance without retrieval augmentation, and the experimental condition will be the same LLMs enhanced with IR-augmented RAG.

Standard LLMs without RAG generate responses based solely on their pre-trained knowledge without access to external documents. Due to their reliance on internal knowledge, which may be incomplete or outdated, we expect these models to have higher hallucination rates and lower factual accuracy.

2 Methods

Our experiment involves constructing a Retrieval-Augmented Generation (RAG) pipeline using LangChain. The system consists of two main components: a retriever, an embedding model, and a generator. The retriever is responsible for extracting relevant documents from an external corpus or database, utilizing vector-based search techniques for efficient similarity matching. For this purpose, we employ FAISS, a library optimized for high-dimensional similarity searches, to store and retrieve document embeddings. The generator is a Gemma 3B model specifically the `gemma3:12b-it-fp16` variant) served through Ollama, which generates responses based on retrieved documents. It is operating with a temperature of 0.1 to ensure deterministic output. Finally, for our embedding model we use the `BAAI/bge-base-en-v1.5` model from Hugging Face, which provides 768-dimensional embeddings optimized for retrieval tasks.

Algorithm 1 rag_basic_idea

```
procedure RAG-GENERATION(question $q$)
    $docs \leftarrow$ similarity_search($q, k = 3$)        ▷ Retrieve top 3 documents
    $context \leftarrow$ clean_retrieved_context($docs$)
    $prompt \leftarrow$ format_prompt($q, context$)
    $answer \leftarrow$ OllamaLLM.invoke($prompt$)
    return $\{answer, context, docs\}$
end procedure
```

We modified the original question-query using a predefined template, which structures the input in a way that enhances retrieval and generation. Given the input query, we first perform information retrieval (IR) to identify the most relevant documents from our database. We use $k = 3$, meaning that the top 3 most relevant documents (retrieved from the FAISS index based on cosine similarity between the query embedding and document embeddings) are retrieved. These documents are then preprocessd and incorporated into a templated query, which structures the input in a way that provides the LLM with additional context. By supplying these retrieved documents along with the modified query, we enhance the model's ability to generate accurate and contextually informed responses while minimizing hallucinations (Figure 1).

2.1 Dataset & Preprocessing

Originally, we used the PolicyQA dataset. However, we found that PolicyQA was not suitable for our needs because many of its questions were based on website privacy policies, which are often repetitive and similar across different sites. This made it difficult to build a meaningful and diverse database for document retrieval, and as a result, our system struggled to return relevant content—leading to weak response quality with RAG.

To address this issue, we switched to the Natural Questions dataset, available on Hugging Face here. This dataset, developed by Google Research, provides a broader and more varied set of real-world questions sourced from web search queries, making it a better fit for our task.

The dataset consists of 307,373 examples in its training set and 7830 in its validation set.

Using Apple's MacBook Pro M1 Max w/64GB ram, we are creating a manageable subset by selecting 300 question-answer pairs from an initial pool of 3,500 samples from the training set and extracting the corresponding documents to build a knowledge base. This ensures clearer questions and more comprehensive reference answers, allowing for a more effective evaluation of the RAG enhancement. Note that we were only able to evaluate 235 of the 300 QA pairs.

The data in the NQ dataset is JSON formatted, and we have provided a python script `nq_processor.py` for parsing. This is used to extract questions, reference answers, and text to create documents for the IR retrieval database. Please see the `prepare_nq_documents` function for how documents are created in the IR retrieval database.

We use huggingface's embedding model BAAI/bge-base-en-v1.5 and create the vector database with FAISS.

2.2 Models

The main model that we tested was `gemma3:12b-it-fp16`, Ollama's Gemma3 with 12B parameters, specifically tuned for following instructions and is a FP-16 full precision unquantized model (retains full accuracy at the cost of space and memory during computation).

We also tested other models such as Deepseek-r1-14b, Deepseek-v2-16b-lite-chat, Gemma3-27b-it-q8, and Qwen2.5-32B-instruct-q5-1

2.3 Evaluation methods

Lexical similarity: To calculate the lexical similarity, we introduced a function to calculate the similarity of two responses using the Jaccard index. It is defined as the size of the intersection of two sets divided by the size of the union of those sets. It is simple, but does not capture semantic similarity as 'car' and 'automobile' would be treated as different words.

Semantic similarity: To calculate the semantic similarity, we calculate the cosine similarities of two responses.

3 Results and Interpretation

Table 1: Comparison of Standard LLM and RAG System

Metric	Standard LLM	RAG System	Difference	Impact
Response Time (s)	37.38	4.56	-32.83	RAG is ∼8x faster
Response Length	1910.43	175.18	-1735.24	RAG responses 90.8% shorter
Semantic Similarity	0.565	0.558	-0.008	Negligible difference (-1.3%)
Lexical Overlap	0.071	0.123	+0.052	RAG has 72.6% better term overlap

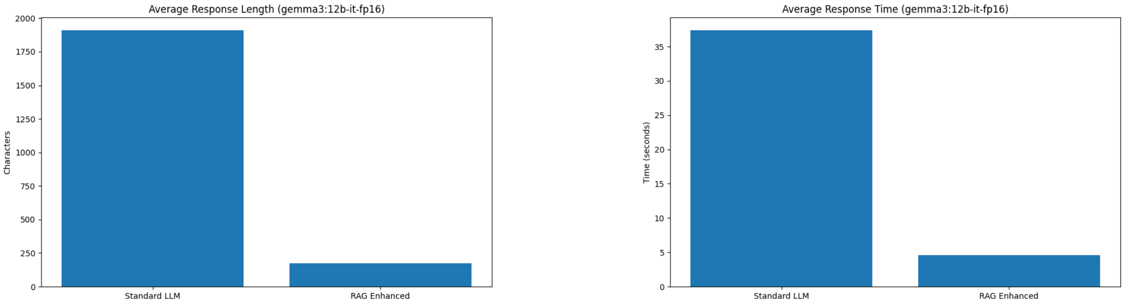

Figure 1: Comparison of average response length (Left), average response time (Right)

3.1 Possible Results and Interpretations

The goal of our project is to reduce hallucinations and improve the accuracy of responses in large language models by leveraging Information Retrieval (IR) and Retrieval-Augmented Generation (RAG). If our hypothesis is confirmed, we expect to see a significant reduction in hallucinations and an improvement in response accuracy when using IR-augmented RAG compared to LLMs without it. Conversely, if our hypothesis is not confirmed, then we expect to see no reduction in hallucinations and no measurable improvement in the accuracy of responses between the two models.

3.2 Results

Initially, our system was built around the PolicyQA dataset and a text generation model, with the goal of evaluating model performance on privacy policy-related questions.The motivation for selecting PolicyQA stemmed from existing research indicating that large language models (LLMs) struggle with accurately interpreting and answering questions derived from privacy policies. We hypothesized that integrating Information Retrieval (IR) and Retrieval-Augmented Generation (RAG) techniques could enhance response accuracy in this domain. However, initial testing using PolicyQA failed to achieve satisfactory results. As discussed in our previous work, we identified several critical limitations: the models consistently failed to produce relevant answers to privacy policy questions, and the incorporation of RAG did not produce any measurable improvement in accuracy. While RAG did demonstrate potential by grounding responses in retrieved content, it became clear that both the choice of dataset and model architecture were constraining its effectiveness. In our revised approach we introduce two key changes to address these issues: (1) the adoption of the Natural Questions dataset, which offers more diverse and general knowledge queries; and (2) the evaluation of performance on chat-optimized LLMs, which are better aligned with conversational QA tasks.

The revised approach yielded positive yet unexpected results. In addition to evaluating accuracy, we assessed several key metrics, including response time, response length, semantic similarity, and lexical overlap, as presented in Table 1. Notably, RAG consistently outperformed the standard model in terms of efficiency, delivering significantly shorter response times and more concise responses, as illustrated in Figures included on Github. On average, RAG achieved an 820% faster response time compared to standard model queries even when accounting for the additional overhead introduced by retrieval. Furthermore, RAG responses were, on average, 90.2% shorter than their standard counterparts, highlighting its efficiency in generating concise answers.These findings support RAG's potential value in automated question-answering systems, particularly in scenarios where speed and brevity are critical.

Beyond performance improvements, RAG also demonstrated improvements in lexical overlap. We observed a 72.6% increase in lexical overlap between the augmented responses and the reference answers, suggesting that RAG incorporated more factually relevant terms from the source documents (refer to Github). In this context, lexical similarity refers to the degree of overlap between the model generated response and the reference answer provided from the Natural Questions Dataset. While lexical similarity does not directly indicate accuracy, a higher lexical overlap suggests that RAG captures key terminology and factual content effectively.

However not all results were favorable. While the conciseness of RAG responses contributed to faster response times, it also raised concerns regarding response completeness. Because the model was explicitly instructed to rely solely on retrieved knowledge, it is possible that information retrieval (IR) did not capture enough relevant information, limiting the model's ability to generate comprehensive responses. Additionally we observed a slight -1.3% decrease in semantic similarity (Figure 3). Semantic similarity measures the degree of overlap in meaning between two pieces of text and is a measure of how similar the underlying meaning of two texts is. A decrease in semantic similarity indicates that RAG did not improve the overall answer quality as measured by this metric.

4 Discussion

4.1 Summary of Results

Our research explored the effectiveness of RAG and IR techniques in reducing hallucinations in Gemma 3 on the Natural Questions dataset. We have produced meaningful results that address our core research questions about the effectiveness of RAG compared to standard LLM responses. Our peers have suggested Comprehensive metrics and visualizations have also been generated to show the clear improvements of each approach (Table 1).

Our findings demonstrate that utilizing Information Retrieval (IR) and Retrieval-Augmented Generation (RAG) into the question-answering system led to significant improvements in response efficiency, with an 820% faster response time and 90.2% shorter response length compared to standard models. Additionally, RAG-enhanced responses exhibited a 72.6% increase in lexical overlap with the reference answer provided from the Natural Questions Dataset. However, despite these improvements, the system faced limitations in completeness, as the conciseness of RAG-generated responses raised concerns about missing contextual details. Moreover, a 1.3% decrease in semantic similarity suggests that RAG did not enhance overall answer quality in terms of underlying meaning. These results highlight the trade-offs between efficiency and response completeness, suggesting that while RAG improves performance, further refinements are needed to ensure comprehensive and semantically rich responses.

4.2 Interpretation of Results

The results suggest that while RAG substantially improves performance in terms of speed and relevance—evidenced by shorter response times, more concise answers, and higher lexical overlap, it does so at the cost of depth and completeness. The increase in lexical overlap indicates that RAG is effective at retrieving and incorporating relevant terms from source documents, which likely improves factual grounding and reduces hallucinations. However, the slight decline in semantic similarity implies that these responses, while faster and more aligned with source text, may omit nuanced or inferential content needed for a fully accurate answer.

However, this may be a result of its strict adherence to retrieved evidence. Unlike the standard LLM which when struggling to provide exact answers, may compensate with speculative/hallucinatory details (Huang et al., 2023). Further human evaluation into each response may be required to determine whether standard LLM responses provided meaningful and correct context/data.

4.3 Improvements & Concerns

Reviews of previous work indicated concerns regarding model choice, evaluation metrics, lack of transparency in document selection, and small-scale sample size with overgeneralized claims. Acknowledging this feedback, we have improved upon existing architecture and changed large portions of our design choices. More specifically, to address these concerns, we have:

1. Transitioned from GPT-2 XL to Gemma-3 (12B-it), fine-tuned for instruction.

2. Implemented automated metrics via. semantic similarity (cosine similarity), lexical overlap (Jacarrad similarity), and response time/length.

3. Largely scaled our sample size by conducted 235 tests (the original 300-target; limited by 64GB RAM).

4. Provided our full retrieval context, data, and report in Appendix A.

Our analysis is now grounded in worthwhile data, as our implementation has significantly improved.

4.4 Future Research Directions

During Retrieval Augmented Generation (RAG), the LLMs were explicitly instructed to rely solely on retrieved knowledge. While this constraint was intended to minimize hallucinations, it also limited the model's ability to infer or incorporate relevant prior knowledge beyond the retrieved content. As a result, responses, though factually grounded, often lacked completeness or contextual depth. A promising direction for future research is to explore methods that strike a balance between strict reliance on retrieved information and the model's capacity for inference.

Another future research direction could be improving the evaluation metrics in order to determine whether the response generated contains a correct answer. Although we evaluate the responses with lexical similarity and semantic similarity, there isn't a hard truth that determines whether the response generated contains an answer that is correct.

4.4.1 Limitations

1. Testing was limited by computational resources, which constrained the full scope of this study. While we aimed to evaluate 300 question-answer pairs locally on our 64GB RAM machine, memory constraints forced us to cap testing at 235 samples instead of the intended 300. Similarly, lab machines tested did not meet hardware requirements for extensive sample sizes.

2. Smaller-scale tests were conducted on alternative models (DeepSeek-V2, DeepSeek-R1, Qwen-2.5, Llama-3) for preliminary insights. However, these models (particularly the Chinese optimized ones) performed poorly on English-based QA tasks. Among the tested models, Gemma-3 proved the most capable, whereas Llama-3 exhibited the most noticeable limitations in factual consistency. A future, more in-depth comparative analysis will be necessary to fully assess each model's RAG capabilities.

4.5 Conclusion

This research demonstrates that RAG system significantly improves speed and factual accuracy by grounding responses in retrieved evidence. It effectively reduces hallucinations, as seen in lengthy, and potentially speculative standard LLM responses.

This work provides important insights for understanding LLMs' limitations in complex domains and points to promising directions for future research.

References

Lewis et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks arXiv:2005.11401v4 Sng et al. Novel Approach to Eliminating Hallucinations in Large Language Model-Assisted Causal Discovery arXiv:2411.12759 Huang et al. A Survey on Hallucination in Large Language Models arXiv:2311.05232 Priola. Addressing Hallucinations with RAG and NMISS in Italian Healthcare LLM Chatbots arXiv:2412.04235

A Appendix

All data files included in Github

Prompt construction follows this template:

```
Please answer the question based on the following
retrieved information. Use only the provided
information and do not use your own knowledge.

Retrieved Information:
[context]

Question: [question]

Answer:
```