INL/MIS-24-79615-Revision-0

Reducing AI RAG Hallucination by Optimizing Routing Techniques

August 2024

Darrin Michael Lea, Rafer Scott Cooley, Michael Adam Cutshaw, Zachary M Priest

Changing the World's Energy Future

INL is a U.S. Department of Energy National Laboratory operated by Battelle Energy Alliance, LLC

INL/MIS-24-79615-Revision-0

Reducing AI RAG Hallucination by Optimizing Routing Techniques

Darrin Michael Lea, Rafer Scott Cooley, Michael Adam Cutshaw, Zachary M Priest

August 2024

Idaho National Laboratory
Idaho Falls, Idaho 83415

http://www.inl.gov

Prepared for the
U.S. Department of Energy
Under DOE Idaho Operations Office
Contract DE-AC07-05ID14517

Reducing AI RAG Hallucination by Optimizing Routing Techniques

Darrin Lea | Louisiana State University | D320 | Rafer Cooley, Michael Cutshaw, Zach Priest

Overview

Large Language Models (LLMs), such as ChatGPT, tend to "hallucinate", meaning they confidently generate false information.

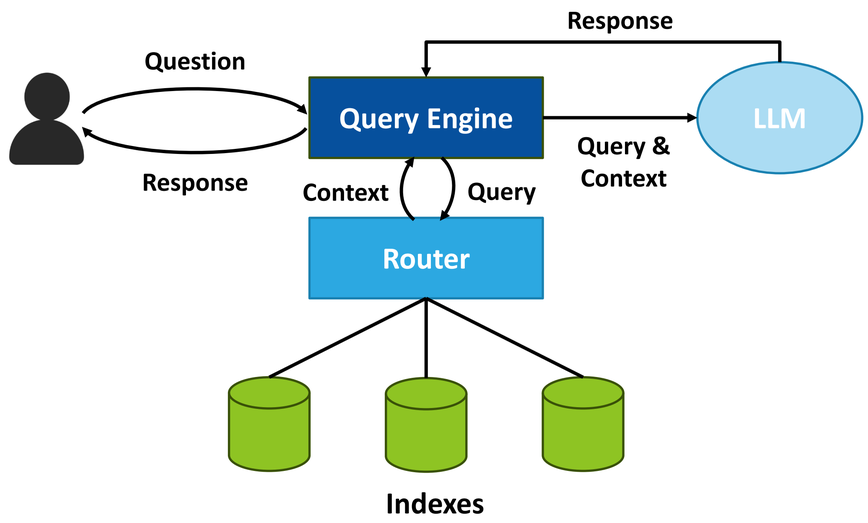
Retrieval Augmented Generation (RAG) attempts to diminish hallucination by providing context to the LLM from data stores (indexes) containing relevant information. The LLM uses this context to formulate its response.

Each index segments information into chunks called nodes by tokenizing the text into a list of numbers, a process called Embedding.

In a RAG system, a Router determines which index the context is retrieved from.

Index Summary Routing is a method where the router analyzes a summary of each index to determine which one contains relevant information.

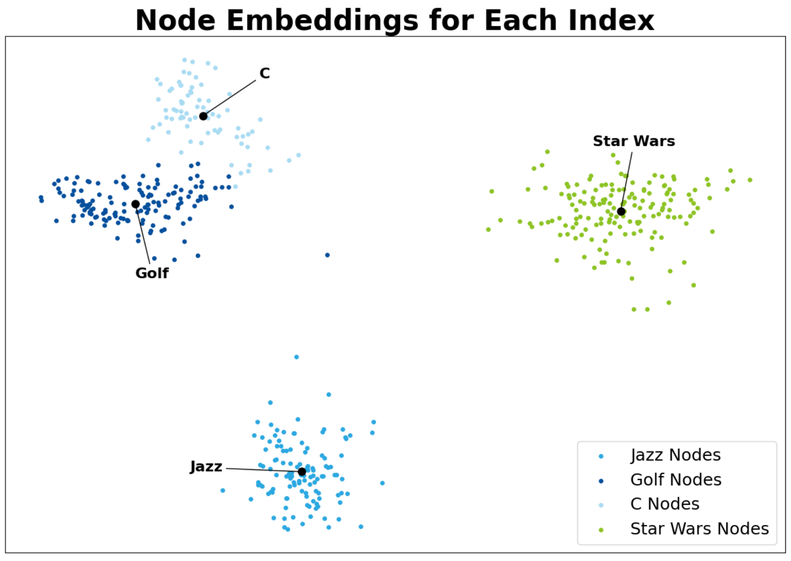
This study aims to optimize Index Summary Routing.



Experimental Design

In this study, we explore different methods of Index Summary Routing to optimize the routing process in a RAG system. We designed the experiment as follows:

1. Create four indexes based on information gathered from the Wikipedia API about these topics – Golf, Jazz, Star Wars, and C programming language
2. Embed the text into nodes, which will be used to generate index summaries. The scatterplot below graphs all nodes for each index as well as the mean embeddings.



3. Generate a summary for each index using the following methods.

Summarization Method	Description
No Summary	No summaries are provided to the router
Manual Summary	Summaries are human-generated
Random Node Summary	Summaries are LLM-generated using random nodes within the index as context
Inner K Node Summary	Summaries are LLM-generated using the K most central nodes as context
Inner & Outer K Node Summary	Summaries are LLM-generated using the K most central and K most remote nodes

4. For each summarization method, query the RAG system with 10 questions from each topic.
5. Record the failure rate of the router for each summarization method.
6. Measure the accuracy and relevance of the answers generated.
7. Perform five rounds of querying for each summarization method.

We used Python's LlamaIndex library, the Ollama library, the HuggingFace transformers library, and the RAGAS framework to conduct these experiments.

Problem

RAG systems still suffer from hallucination because of bad embeddings or bad routing. For example, a router will often return context from an irrelevant index, resulting in a hallucinated answer.

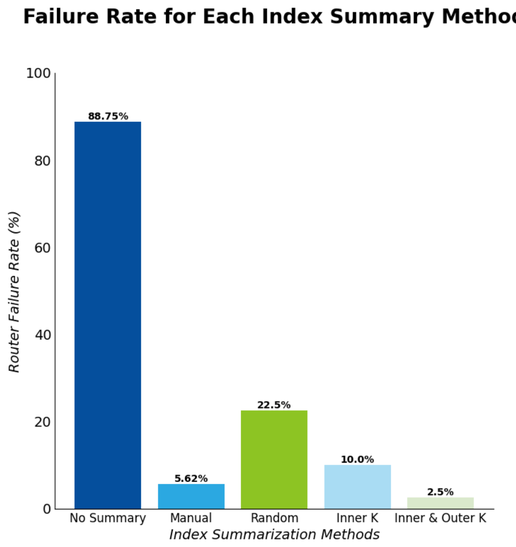
In this study, we aim to minimize the frequency of routing hallucinations by optimizing Index Summary Routing.

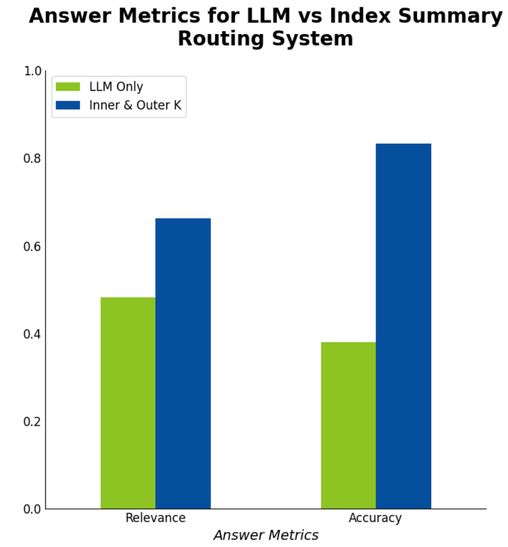
Results

The left chart compares the summary methods based on their failure rate (the percentage of questions where the router chose the wrong index)

- Besides No Summary, Random Node summarization yielded the highest error rate, likely because it contains very few of the index's central nodes.
- The Inner & Outer K summarization method proved the most effective.

The right chart shows answer metrics for a standalone LLM alongside a RAG architecture using the Inner & Outer K Index Summarization Method. There is an increase in answer accuracy and relevancy (decrease in hallucinations) when using the RAG Inner & Outer K architecture.





Conclusion

Overall, we found the Inner & Outer K method to be the optimal summary strategy for a router in a RAG system. This is likely because it provides the most comprehensive overview of the index, containing the most relevant information along with outlier information.

In the future, we would like to explore the possibility of auto-indexing a data source, such as the entirety of Wikipedia, using the optimal summary method.