s.103053

09 december 2024

QUANTITATIVE RISK ASSESSMENET OF LLMS IN BUSINESS

Evaluating Retrieval-Augmented Generation Using the EU AI Act

Quantitative Risk Management – An Application of Machine Learning

Individual Exam Paper

29,500 Characters = 13 pages

CBS COPENHAGEN BUSINESS SCHOOL
HANDELSHØJSKOLEN

TABLE OF CONTENTS

1. INTRODUCTION

"AI is a tool. The choice about how it gets deployed is ours." – Oren Etzioni

In the evolving landscape of artificial intelligence (AI), businesses increasingly adopt large language models (LLMs) for generating and summarizing text (Fu, Hadid, & Damer, 2024). However, these powerful systems are not without limitations. One significant challenge is the occurrence of "hallucinations," where the models generate responses that, while syntactically plausible, are factually incorrect (de Prado, 2018). This is particularly problematic in business and legal contexts, where precision and accuracy are paramount to avoid compliance risks, reputational harm, or incorrect interpretations of critical documents (Buchicchio, et al., 2024).

To address these concerns, Retrieval-Augmented Generation (RAG) has emerged as a promising solution. By integrating external, up-to-date data sources into the LLM's response generation, RAG reduces reliance on the model's static pretraining knowledge, thereby enhancing factual accuracy with accurate context (Wiratunga, 2024). This report evaluates the quantitative risks of using LLMs in a business context and how implementing RAG can help to mediate these risks.

The study uses the EU AI Act, a newer legal document/framework (EU Artificial Intelligence Act, 2024), as the dataset and Llama3.1 as the LLM. Released after the training data cutoff for Llama3.1, the EU AI Act presents an ideal case to assess the efficacy of RAG in mitigating hallucination risks. By comparing the model's performance with and without RAG when interpreting query's regarding the Act, this research aims to quantify improvements in the LLMs responses by reducing error rates, contributing to the broader understanding of LLM applications in high-risk applications.

This report begins by outlining the risks and benefits of using LLMs in business settings and introduces the specific challenges addressed in this study. It then details the experimental setup, including the metrics for evaluating the model's performance, before presenting the results of the quantitative analyses that compare the error rates and reliability of Llama3.1 with and without RAG. The findings are then further explored to discuss their implications, the study's limitations, and potential areas for improvement. Finally, the report concludes with key insights and recommendations for the safe and effective deployment of LLMs in business contexts.

By addressing these critical aspects, this report seeks to contribute actionable insights into the integration of AI technologies in risk-sensitive domains such as law and compliance.

2. PROBLEM STATEMENT

Quantitative risk assessment is a cornerstone for modern businesses to manage uncertainty, identify vulnerabilities, and allocate resources effectively (de Prado, 2018). The subject highlights not only the importance of assessing risks in business contexts but also the risks inherent in the quantitative models themselves. While these models are powerful, they present challenges such as interpretability, overfitting, and governance (de Prado, 2018). Large Language Models (LLMs), although perceived as natural language processors by the average user, are fundamentally complex statistical systems that generate systematic and quantitative outputs. As such, their implementation and use are closely tied to the principles of quantitative risk assessment (Buchicchio, et al., 2024).

Quantitative models play a critical role in identifying, measuring, and mitigating risks in complex environments (de Prado, 2018). They allow organizations to evaluate scenarios of uncertainty by offering structured methods to calculate probabilities and assess the potential impacts of adverse events. According to Hansson's philosophical analysis, quantitative risk models strike a balance between certainty and epistemic uncertainty, enabling organizations to translate complex uncertainties into actionable probabilities (Hansson, 2004). This capability is especially vital in fields where decisions hinge on probabilistic outcomes, such as financial services or technology deployment (Liebergen, 2024).

Despite their benefits, quantitative models are not without risks. Their effectiveness depends on data integrity, the validity of underlying assumptions, and the potential for overfitting, all of which can introduce vulnerabilities (de Prado, 2018). As Hansson emphasizes, while Bayesian decision theory provides a framework for managing epistemic uncertainty, its limitations underscore the cognitive boundaries of human decision-makers (Hansson, 2004). Additionally, the complexity of machine learning models, often labelled as "black boxes," complicates interpretability and governance, as Liebergen observes in his analysis of AI applications in financial systems (Liebergen, 2024).

Similarly, LLMs are prone to hallucinations—where the model generates statistically plausible but factually incorrect responses, often because it lacks the necessary contextual metadata and because questions are on data that the model was not trained on (Buchicchio, et al., 2024). Retrieval-Augmented Generation (RAG) aims to address this issue by providing external context, enabling the model to deliver factually accurate reasoning.

This study focuses on examining the application of Large Language Models (LLMs) in business contexts, with a particular emphasis on their use in the legal domain. The legal sector was chosen because it has seen significant adoption of natural language models, especially as a tool for text summarization, law cross-validation and interpretation (Buchicchio, et al., 2024). LLMs provide legal professionals with significant time efficiency gains by automating the interpretation of complex documents and rapidly generating summaries or analyses. These tools reduce the manual workload associated with reviewing extensive regulatory texts, enabling professionals to focus on strategic tasks that require human judgment (Buchicchio, et al., 2024).

However, while LLMs can be especially powerful tools for this purpose due to their speed and systematic approach, they also pose significant risks from LLM hallucinations (Buchicchio, et al., 2024). These hallucinations are a substantial risk, as they can be challenging to detect and may lead to serious consequences, such as regulatory non-compliance. To mitigate this risk, the models must be deployed with strong guardrails and risk-mitigating tools such as RAG. This study seeks to utilize and examine RAG as a tool to mitigate risk. By leveraging RAG-enhanced LLMs the goal is to reduce

hallucinations and the models error rating while also showing, that with proper LLM governance, these models have great potential to be powerful tools for quantifying text data. In pursuing this, the study aligns closely with the objectives of the course.

2.1. OBJECTIVE:

This study aims to evaluate the effectiveness of Retrieval-Augmented Generation (RAG) in improving the accuracy of LLMs using a pre-trained open-source model, Llama3.1. The study focuses on queries regarding the EU AI Act—an emerging regulatory framework introduced after the model's data cutoff. By comparing Llama3.1 responses with and without RAG, the project aims to quantify the impact of RAG retrieval on the model's error rates and hallucination tendencies, providing insights into the benefits and limitations of RAG integration.

3. METHODOLOGY

The objective of this study is to assess how LLMs, such as Llama3.1, can be implemented with RAG to mitigate risks associated with hallucinations in high-risk scenarios such as reading legal documents. To evaluate the objective of this study, we will first be deploying a local LLM to ensure that the model used is pretrained, without RAG and without internet access. This pretrained model will then be asked selected questions that require having read the EU-AI act to demonstrate how a LLM might hallucinate when queried about information it wasn't trained upon.
Hereafter, the RAG system must be built as an additional layer in the query system before the model again is asked the same questions. Now able to retrieve contextual data about the document in question, the model should, in theory, be able to provide much more precise responses grounded in contextual truth. These RAG-enabled responses will then be compared to the responses without RAG, and using cosine similarity, an error rating for both responses will be quantified for comparison.

3.1. MODEL AND DATA DESCRIPTION

This study employs Metas open source Llama3.1 with 8B parameters as the primary LLM. A complete overview of the model can be seen here:

	Training Data	Params	Input modalities	Output modalities	Context length	GQA	Token count	Knowledge cutoff
Llama 3.1 (text only)	A new mix of publicly available online data.	8B	Multilingual Text	Multilingual Text and code	128k	Yes	15T+	December 2023
		70B	Multilingual Text	Multilingual Text and code	128k	Yes		
		405B	Multilingual Text	Multilingual Text and code	128k	Yes		

Llama3.1 with 8B parameters was chosen due to its advanced applicability, smaller size and diverse training data, meaning that it was applicable in a wide array of tasks. Moreover, because its knowledge cutoff point is December 2023, it's trained data did not include the EU-AI-act, which was published in July 2024, thereby making questions regarding this legal document, an excellent case study for showing how RAG can address and reduce the risk of hallucinations.

The model was downloaded via the Huggingface platform and deployed on a Ucloud cloud computing centre for testing and results generation. The model is free and open sourced; however, one needs to apply for the use of the model via either Huggingface or Metas Llama webpage. Once access has been granted, the model is available for deployment. Huggingface was chosen as the primary runner of the model, since it provides an excellent platform for model pipeline deployment with a high level of ease of use. Still, a considerable amount of time was dedicated to the development and result generation of the model which is available either in the attached .zip file or on the provided GitHub link. To save space and keep the report concise, a detailed explanation of how the code works will

not be provided here. A Jupiter notebook of how the code works can likewise be found attached to this report or on the provided GitHub link in the appendix.

The EU-AI-Act was chosen as the primary data document because of its release and nature – a significant legal regulatory document released after the model's knowledge cutoff point – but also because of its obvious relation to the cases study of the risks involved with the use of LLMs and Generative AI. There is a strange satisfaction in having the AI model read its own regulations and provide responses on this document.

3.2. RETRIEVAL-AUGMENTED GENERATION (RAG)

Retrieval-Augmented Generation (RAG) is a hybrid approach that combines the strengths of Large Language Models (LLMs) with an external retrieval system to improve the accuracy and relevance of generated responses (Huang & Wang, 2024; Wiratunga, 2024). Unlike traditional LLMs, which rely solely on pre-trained knowledge, RAG incorporates information retrieved from external sources, such as a document repository or knowledge base, in real-time (Wiratunga, 2024). This enables the model to dynamically supplement its outputs with up-to-date and domain-specific information, significantly enhancing its reliability and reducing its dependence on potentially outdated training data.

The RAG mechanism involves two primary components: a retriever module and a generative module. The retriever module searches a pre-defined repository of documents or data to identify information that is most relevant to the input query (Huang & Wang, 2024). This process often employs advanced similarity search techniques, such as embedding-based matching, to rank passages or documents by relevance. Once the retriever identifies the relevant data, the generative module combines this information with the LLM's pre-trained capabilities to produce a response. By grounding its output in the retrieved data, the model ensures that the generated text is both contextually accurate and well-supported.

The use of RAG is particularly advantageous in high-risk applications, such as legal document analysis, where precision and factual grounding are essential (Buchicchio, et al., 2024). One of its most significant benefits is its ability to reduce hallucinations, a common issue with LLMs where models generate plausible but incorrect information. By anchoring responses to verified external sources, RAG minimizes this risk and ensures outputs are more reliable (Wiratunga, 2024). Furthermore, RAG addresses the limitations of fixed training data cutoffs in LLMs. For example, in this study, the model's training ended in December 2023, but by using RAG, the model dynamically retrieves data from the EU AI Act, which was published in July 2024. This allows the model to incorporate up-to-date legal information that would otherwise be inaccessible.

Another critical advantage of RAG is its adaptability across domains. Without the need for retraining, the model can specialize in specific fields by leveraging a tailored retrieval database (Huang & Wang, 2024). This modular design not only simplifies domain adaptation but also reduces computational costs and improves scalability. In this study, the integration of RAG with Llama3.1 ensures context-based interpretation of the EU-AI act, effectively bridging the gap between the model's pre-trained knowledge and real-time domain-specific requirements. By doing so, RAG plays a central role in enhancing performance and mitigating risks, aligning closely with the principles of robust risk management in AI applications (Liebergen, 2024).

3.3. METRICS AND MODEL EVALUATION

To define and evaluate the model's performance, the key metric chosen here is Error Rate (ER). The error rate quantifies the sematic deviation of our model-generated outputs (legal interpretations or summaries) from the ground-truth responses (short valid correct answers/interpertations). In it's simplest form, it is just 1 – Accuracy (Deep AI, 2024). In our case, accuracy is calculated using cosine similarity between the embeddings of the model's outputs and the ground truth and then subtracting that similarity from 1.

Our formular for calculating the Error Rate is therefore:

$$Error\ Rate(ER) = 1 - COSINE\ Similarity$$

Where Cosine Similarity measures the sematic closeness between two pieces of text (embedding vectors) and ranges from 0 (no similarity) to 1 (perfect sematic match) (Karabiber, 2024). Error Rate therefore ranges from 0 (no error; perfect match) to 1 (maximum error; complete misalignment).

Cosine Similarity is a metric that quantifies the similarity between two vectors by measuring the cosine of the angle between them (Karabiber, 2024). It focuses on the vectors' direction or orientation while ignoring their magnitude or scale. For this calculation, both vectors must belong to the same inner product space, ensuring they produce a scalar value when multiplied using the inner product operation. Smaller angles between vectors produce larger cosine values, indicating greater cosine similarity (Karabiber, 2024).

The formula for cosine similarity:

$$similarity(A, B) = \cos(\theta) = \frac{A \cdot B}{||A||\ ||B||}$$

3.4. WORKFLOW FOR CALCULATING ERROR RATE

To create the dataset, 20 questions were generated from the EU-AI act and a short, interpreted ground truth on these questions was defined. These questions were then presented to the model as query's, with and without rag and added to the dataset. To calculate the error rate and cosine similarity on this data, we must transform the text into numerical vector representations, also called text embeddings. Using a pre-trained sentence transformer model, these text embeddings were created for both the true answers, the no RAG answers, and the RAG answers. Finaly, the cosine similarity between the ground truth embeddings and the model-generated ouputs are computed and the error rate is found:

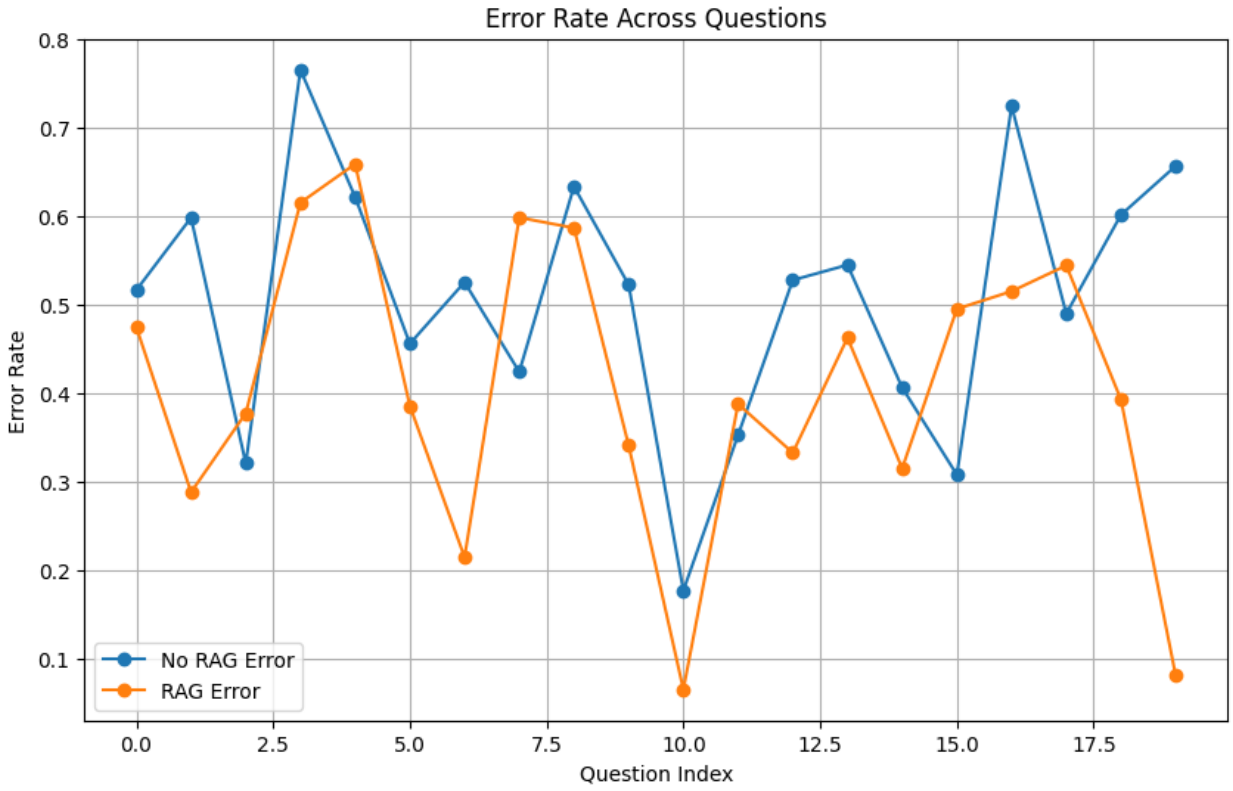
$$\cos(\theta) = \frac{True\ Embedding \cdot Model\ Embedding}{||True\ Embedding||\ ||Model\ Embedding||}$$

$$Error\ Rate = 1 - \cos(\theta)$$

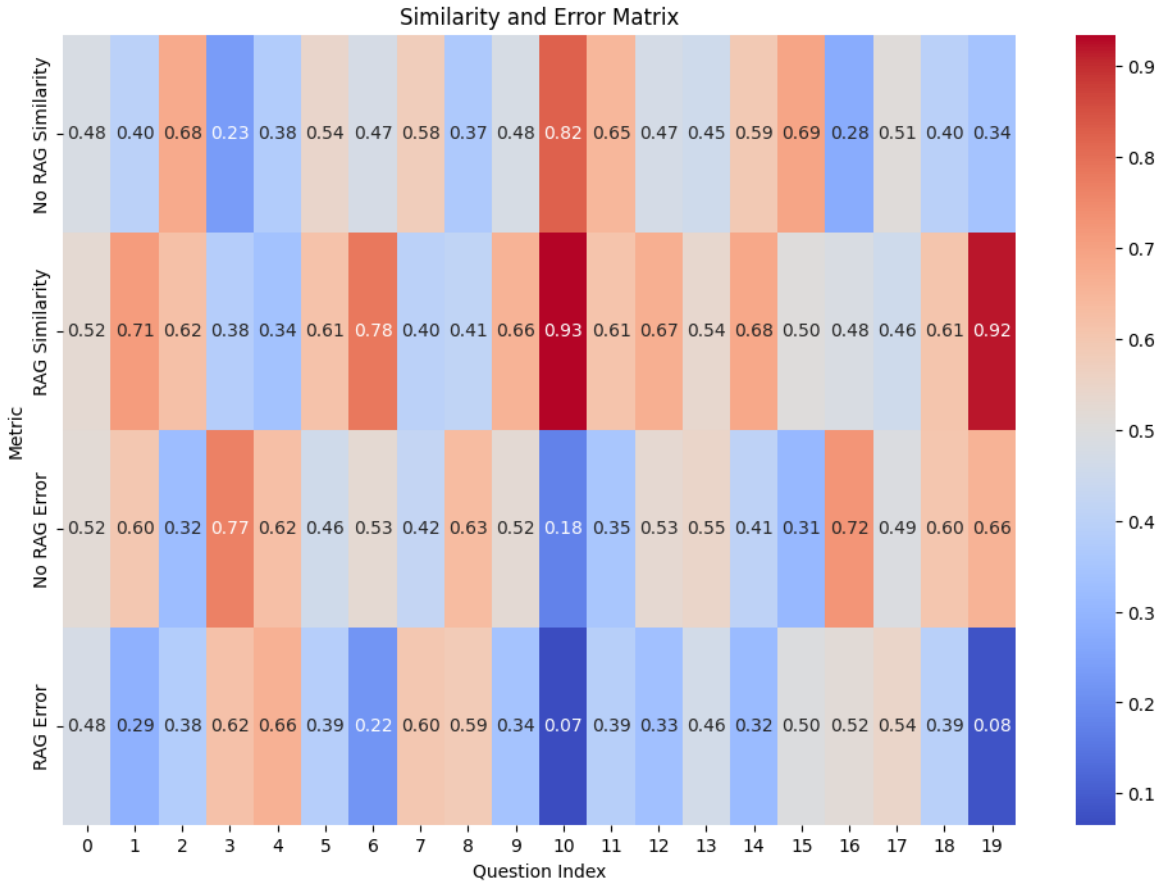
These calculated error rates are then inserted into our dataset and available for us to compare and visualize as the results of this study. A complete overview of how this data was coded, calculated and visualized can be found in the *ER_Calculations.ipynb* file attached in the .zip or online at GitHub.

4. RESULTS AND FINDINGS

In this section of the study the results will be presented and examined. Looking at the results, the comparison between Llama3.1 with RAG and without RAG demonstrates a moderate yet clear advantage of RAG across the similarity score and error rate in most cases.



As observed in the line chart above, having RAG outperforms the baseline No-RAG model on most observations, with significant differences at indices 1, 6, and 19. For instance, question 19, *"What is the purpose of the EU AI regulation act?"* shows a major difference between the baseline No-RAG models response and the model with RAG, illustrating the models ability to retrieve correct context from the data and provide a clear response to the question. A heatmap was also generated to provide another overview of the exact score of each instance.



The heatmap further provides a granular breakdown of performance across all indices, illustrating how RAG mitigates key risks such as hallucinations. Without RAG, the Llama3.1 model frequently produces higher error rates, such as indices 1, 3, 16, and 19, where limited contextual information misled the model into hallucinations and incorrect responses. These instances show that the model will produce a response but because it has not been trained nor given the contextual data from the EU-AI act, it will produce a response based on hallucinations. In contrast to this, the responses where RAG is applied, produce a response that is grounded in contextual information retrieved from the data and therefore is closer to the truth.

Despite its overall robustness, the RAG-enhanced model exhibited some limitations where contextual data was insufficient or misaligned with the query. For example, 0, 7, and 17 show both models struggling to produce responses similar to the ground truth answer. This indicates that even with RAG, the model faced challenges in providing accurate responses to perhaps highly nuanced or ambiguous questions, or that perhaps the correct contextual data was not provided efficiently.

Overall, the results indicate that RAG enhances model performance by reducing the overall error rates. Calculating the average error rate for the No-RAG model and the RAG model, we find the NO-RAG average error rate to be approximately 0.51 which is to be expected when most of the questions are of an interpretative nature. The average error rate for the RAG model is slightly lower at approximately 0.41 which indicates that these improvements, while smaller than expected, are still significant and demonstrates that RAG does serve to mitigate the risk of hallucinations when applying LLMs to texts. These findings will be further discussed in the subsequent section, where we address the model's limitations and outline potential areas for refinement.
For further illustrations of the findings, please see the appendix, the *ER_Calculations.ipynb* file or online at GitHub link.

5. DISCUSSION

While implementing and evaluating the Retrieval-Augmented Generation (RAG) system, several limitations and areas for improvement were identified within the codebase and methodology. These observations provide insight into why certain outputs may not have fully met expectations and highlight opportunities for future refinement.

5.1. RETRIEVAL SYSTEM AND QUERY LIMITATIONS

One key discovery pertained to the retrieval system's default configuration, which limited the model to retrieving a maximum of five chunks of context from the embedded text stored in the Chroma database. This constraint meant that for queries requiring comprehension of broader sections of the EU AI Act, the model's output could not encompass the complete truth, as relevant contextual information was omitted. This limitation significantly impacts the model's ability to handle queries that demand analysis of intricate or interdependent sections of the document, potentially leading to incomplete or misleading responses.

Incomplete or misleading responses are some of the most significant risks when using LLMs in such cases (Buchicchio, et al., 2024). The baseline model, in particular, frequently generated incorrect or hallucinated responses, showcasing the inherent dangers of deploying models without appropriate safeguards. In practical applications, such as legal or regulatory interpretations, misinterpretations of critical documents like the EU AI Act could result in flawed decision-making, non-compliance, or reputational damage for organizations (Wiratunga, 2024). Increasing the retrieval limit and exploring dynamic retrieval techniques based on query complexity would likely enhance the system's overall accuracy, mitigating the risks associated with applied LLMs.

Another area identified for improvement involves formulating better questions and providing a more expansive "whole-truth" dataset against which the model's outputs are compared. The current setup, where the ground-truth data reflects interpretations or simplifications of the EU AI Act, introduces an additional layer of subjectivity into the evaluation process. By improving question design to eliminate ambiguity and ensuring that the ground-truth data includes a broader and more detailed representation of the source material, the model's performance can be evaluated more reliably. For example, Question 19 demonstrated how well the model performed when provided with clearer framing and a more encompassing ground truth based on actual quotes from the EU-AI-act.

If this study is to be redone or continued, a comprehensive rework of the query framework should be done, where each question is given a much more detailed and comprehensive ground truth to compare to. Consequently, if the model is wrongly evaluated or its limitations are overlooked, organizations may overestimate its reliability, leading to improper usage or decision-making processes (Buchicchio, et al., 2024). These limitations underline the need for enhanced evaluation frameworks, which are discussed in the following section.

5.2. EXPANDING EVALUATION METRICS

The evaluation of the Llama3.1 model in this study focused on error rates, which, while valuable for assessing hallucinations and contextual accuracy, do not fully capture the model's effectiveness. Inspired by Buchicchio et al. (2024), *Design, Validation, and Risk Assessment of LLM-Based Generative AI Systems Operating in the Legal Sector*, additional metrics could offer a more nuanced understanding of performance. Text quality and content coherence, measured through human assessments like Likert-scale questionnaires, can evaluate clarity, conciseness, and alignment with the source. Metrics such as the Text Quality Score (QTXT) highlight these aspects and could provide qualitative comparisons between RAG outputs and baseline or ground-truth answers.

The model's ability to handle multiple topics is another critical factor. Metrics like the Multiple Topics Quality Score (QMULTI) assess whether responses address complex or overlapping queries comprehensively and accurately (Buchicchio, et al., 2024). Similarly, evaluating keyword extraction accuracy with a Keyword Quality Score (QKEY) can measure how well the model identifies and prioritizes essential concepts, which is particularly important for context-heavy tasks like regulatory interpretation. By combining these qualitative metrics with error rate analysis would provide a more holistic evaluation, offering deeper insights into the model's strengths and limitations in real-world applications, which ultimately will help limit the risks involved in implementing such interpretative LLMs in businesses (Buchicchio, et al., 2024).

While this study focuses primarily on the Retrieval-Augmented Generation (RAG) approach, it is important to contextualize its performance relative to other modelling techniques for mitigating risks in LLM outputs. Fine-tuning an LLM, such as Llama3.1, on domain-specific data represents one viable alternative. By retraining the model on a curated dataset of legal documents and regulatory texts, it is possible to enhance its ability to interpret the EU AI Act without being solely dependent on real-time retrieval systems. This method offers deeper contextual understanding and can improve the model's reliability for repeated or domain-specific queries, which is especially useful if the model needs to be deployed on a wider array of legal documents. However, fine-tuning requires significant computational resources and access to high-quality annotated data. Furthermore, fine-tuned models may suffer from overfitting, limiting their ability to handle diverse or unforeseen queries outside their training domain. Evaluation metrics alone, however, cannot address all the risks of deploying LLMs. Governance and oversight remain crucial, as detailed in the next section.

5.3. GOVERNANCE AND RISK MANAGEMENT CHALLENGES

Effective governance and risk management are still essential to ensure the responsible deployment of LLMs in business and regulatory settings (Liebergen, 2024). Given the high stakes associated with interpreting documents like the EU AI Act, governance frameworks are needed to ensure transparency, accountability, and robustness in AI systems (Buchicchio, et al., 2024). The EU AI Act even categorizes LLMs deployed for judicial or regulatory interpretations as high-risk AI systems, mandating strict requirements for transparency, bias mitigation, and human oversight (EU Artificial Intelligence Act, 2024).

A lack of sufficient governance exposes organizations to several risks. For instance, bias amplification can occur if training or retrieval data contain unrecognized biases, leading to systematic errors in AI outputs (Buchicchio, et al., 2024; Leo, Sharma, & Maddulety, 2019). This was evident in this studies model, where if only sections of the required context were presented to model via the RAG retriever, then the model would not present a wholistic or accurate answer. Similarly, poor data handling practices may compromise sensitive information, increasing compliance and security risks (Leo, Sharma, & Maddulety, 2019). Another critical challenge is the lack of explainability in LLM outputs. As these models often function as black-box systems, their inability to provide clear, interpretable

reasoning for their outputs can hinder validation efforts and erode trust in AI-driven decisions (Fu, Hadid, & Damer, 2024).

To address these governance challenges, organizations must establish comprehensive frameworks tailored to AI systems. These frameworks should include policies for continuous monitoring of outputs to identify inaccuracies, biases, or data inconsistencies in real-time (Buchicchio, et al., 2024). Additionally, integrating human oversight remains essential to ensure that AI outputs align with organizational goals, ethical standards, and regulatory requirements (EU Artificial Intelligence Act, 2024). By combining automated AI capabilities with expert validation, businesses can mitigate risks associated with bias, misinterpretation, or poor data handling.

5.4. IMPROVEMENTS TO CODEBASE

It is also crucial to acknowledge that the codebase was developed primarily to validate a specific hypothesis—that RAG reduces the Error Rate of LLMs when applied to unseen data. As such, the implementation prioritizes simpler proof of concept over highly accurate optimized performance. Several aspects of the system could be enhanced, including refining prompt engineering techniques to better align with the model's strengths and weaknesses, and improving retrieval configurations to dynamically adjust context retrieval based on the query's complexity.

Finally, the choice of underlying models and embeddings also offers room for improvement. While the current setup demonstrates the utility of RAG, employing more sophisticated models or embeddings could yield greater gains in performance. For instance, leveraging more advanced text embeddings or adopting retrieval strategies that integrate with larger or domain-specific models might enhance the system's ability to capture nuance and deliver more accurate outputs.

These observations underscore that while the current system effectively demonstrates the advantages of RAG, significant potential remains for further development and optimization. Addressing the identified limitations and exploring these avenues for improvement would not only enhance performance but also strengthen the system's applicability to complex, high-stakes domains like regulatory compliance.

6. CONCLUSION

This study evaluated the application of Retrieval-Augmented Generation (RAG) in improving the accuracy and reliability of Large Language Models (LLMs) for high-stakes tasks, particularly in the interpretation of complex legal documents such as the EU AI Act. By comparing the performance of the Llama3.1 model with and without RAG, the analysis revealed meaningful insights into the strengths and limitations of this approach.

The integration of RAG into Llama3.1 led to a measurable reduction in error rates, lowering the average from 0.51 to 0.41. This improvement demonstrates RAG's capability to mitigate the risks of hallucinations by grounding model responses in relevant external data. The RAG-enhanced model outperformed the baseline model in providing accurate and contextually grounded answers, particularly for straightforward and direct queries. However, both setups faced challenges with ambiguous or nuanced questions, often due to limited retrieval configurations or insufficient context in the embedded data. These results underscore that while RAG can significantly improve LLM outputs, it is not a comprehensive solution and cannot entirely eliminate inaccuracies.

For businesses, the findings highlight both the promise and the limitations of deploying LLMs in risk-sensitive contexts, such as regulatory compliance, legal analysis, or decision-making frameworks. The ability of RAG to reduce hallucinations makes it a valuable tool for augmenting human expertise, particularly in tasks that require analysing new or sensitive documents. However, reliance on these systems without appropriate safeguards can expose organizations to substantial risks, including non-compliance, flawed decisions, or reputational damage. Effective governance frameworks are crucial to mitigate these risks, emphasizing transparency, accountability, and human oversight. As outlined in EU-AI-act, organizations must treat LLM outputs as an augmentation of expert judgment rather than a replacement, ensuring that automation aligns with organizational goals and ethical standards.

Future research and development should prioritize enhancing the evaluation metrics used to assess LLM performance. While error rate provides valuable insights into accuracy, additional metrics such as text quality, coherence, and topic coverage would offer a more nuanced understanding of model effectiveness. Incorporating advanced metrics like the Multiple Topics Quality Score or the Keyword Quality Score could provide deeper insights into the model’s strengths and areas for improvement.

In conclusion, this study illustrates the value of RAG in enhancing the reliability of LLMs for critical tasks while highlighting the importance of addressing inherent limitations. As businesses increasingly adopt AI-driven solutions, balancing innovation with risk awareness will be an essential and ongoing process. Continuous research, improved governance, and technical optimization will ensure that LLMs can safely and effectively support decision-making in complex and high-stakes environments.

7. REFERENCES

Buchicchio, E., Angelis, A. D., Moschitta, A., Santoni, F., Marco, L. S., & Paolo, C. (2024). Design, Validation, and Risk Assessment of LLM-Based Generative AI Systems Operating in the Legal Sector. *International Symposium on Systems Engineering (ISSE)*.

de Prado, M. L. (2018). *Advances in Financial Machine Learning.*

Deep AI. (2024, December 10). *Deep AI*. Retrieved from Accuracy (error rate): https://deepai.org/machine-learning-glossary-and-terms/accuracy-error-rate

EU Artificial Intelligence Act. (2024, July). *EU Artificial Intelligence Act.* Retrieved from The Act Texts: https://artificialintelligenceact.eu/the-act/

Fu, B., Hadid, A., & Damer, N. (2024, November 30). Generative AI in the context of assistive technologies: Trends, limitations and future directions. *Image and Vision Computing*.

Hansson, S. O. (2004). *Philosophical Perspectives on Risk.* Stockholm: Royal Institute of Technology.

Huang, D., & Wang, Z. (2024). *Evaluation of Orca 2 Against Other LLMs for Retrieval Augmented Generation.* Singapore: School of Computing and Information Systems, Singapore Managment University.

Karabiber, F. (2024, December 3). *Cosine Similarity*. Retrieved from LearnDataSci: https://www.learndatasci.com/glossary/cosine-similarity/

Leo, M., Sharma, S., & Maddulety, K. (2019, March 5). Machine Learning in Banking Risk Management: A Litterature Review. *MDPI*.

Liebergen, B. v. (2024). Machine Learning: A Revolution in Risk Managment and Compliance? *The Capco Institute Journal of Financial Transformation*.

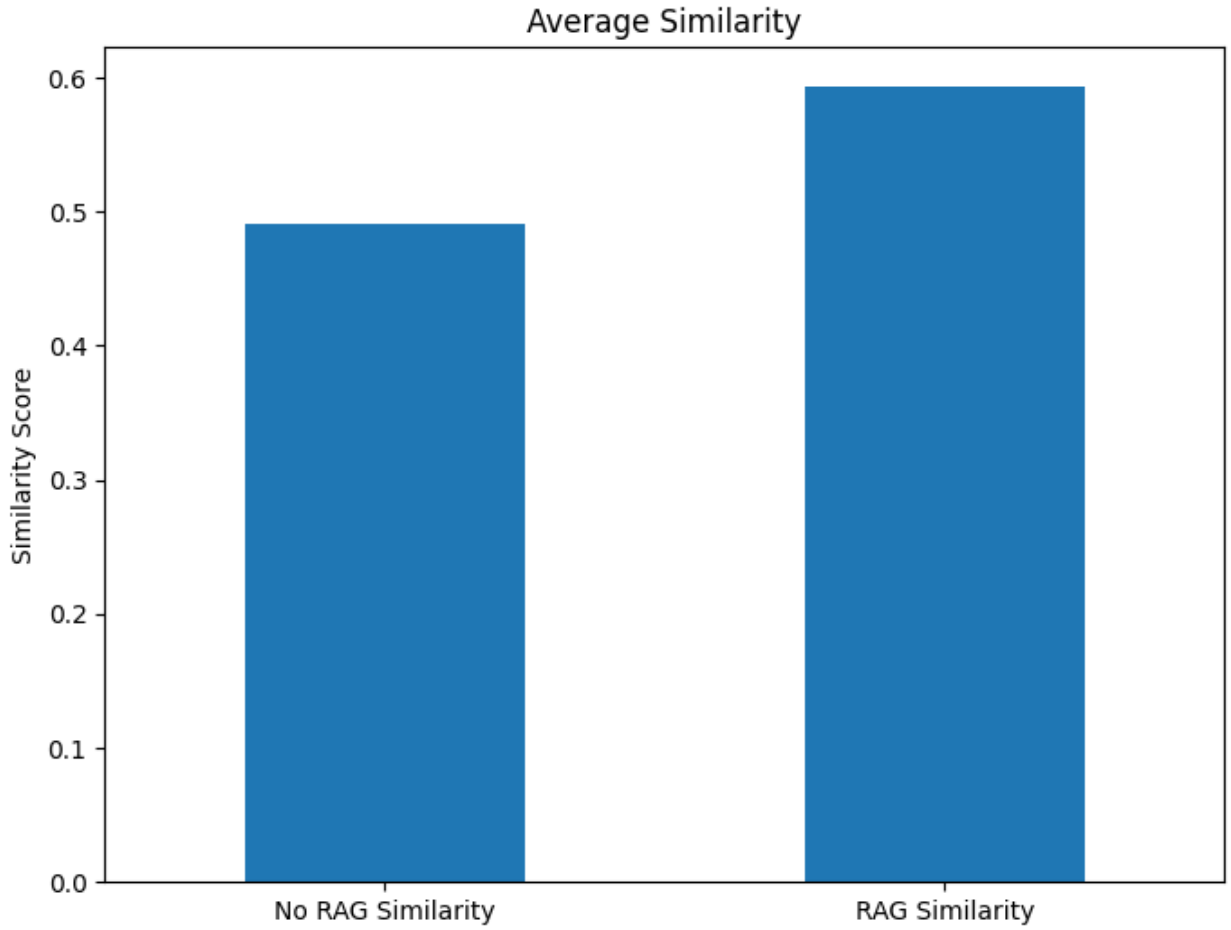
Wiratunga, N. e. (2024). *CBR-RAG: Case-Based Reasoning for Retrieval Augmented Generation in LLMs for Legal Question Answering.* Switzerland: International Science Partnerships Fund.

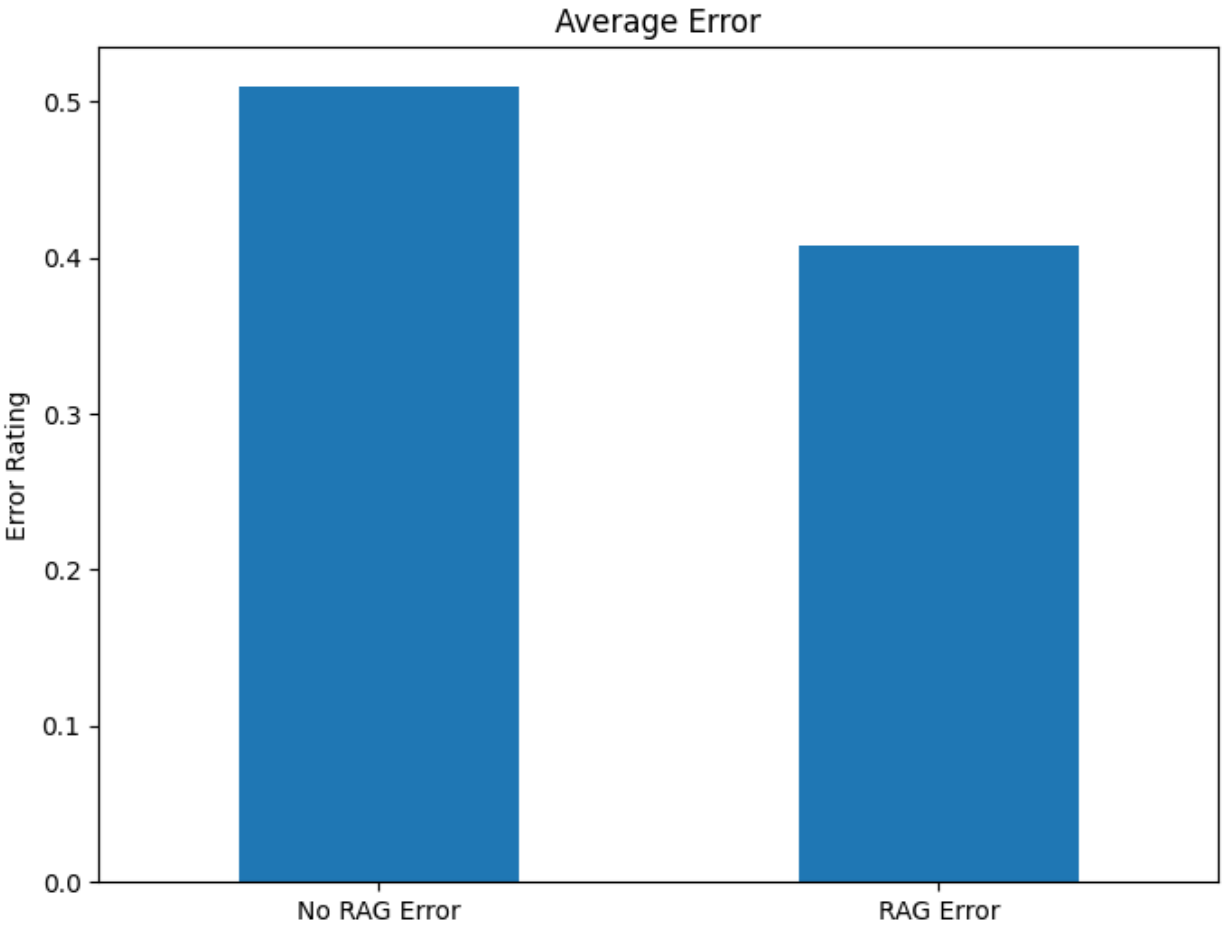
8. APPENDIX

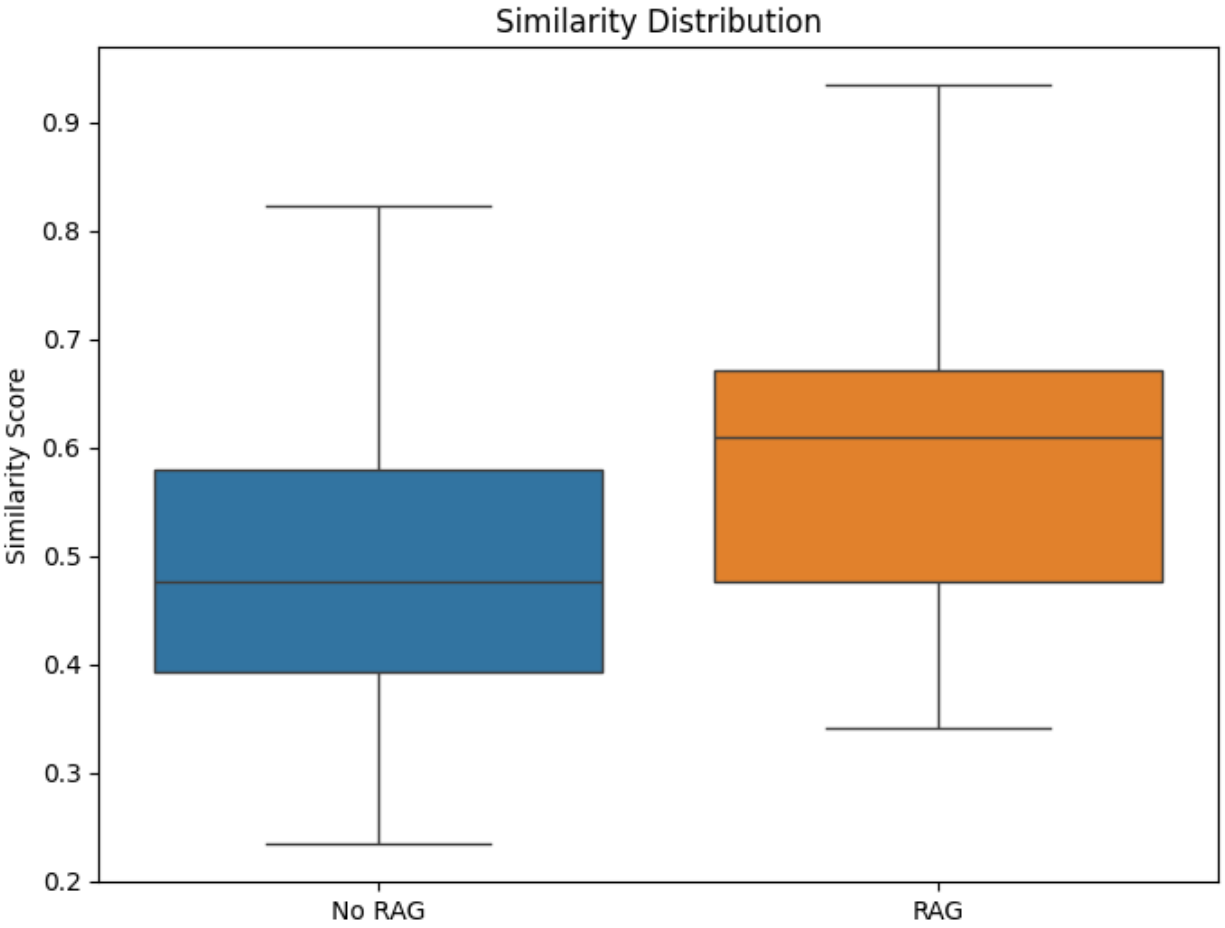
8.1. GITHUB LINK

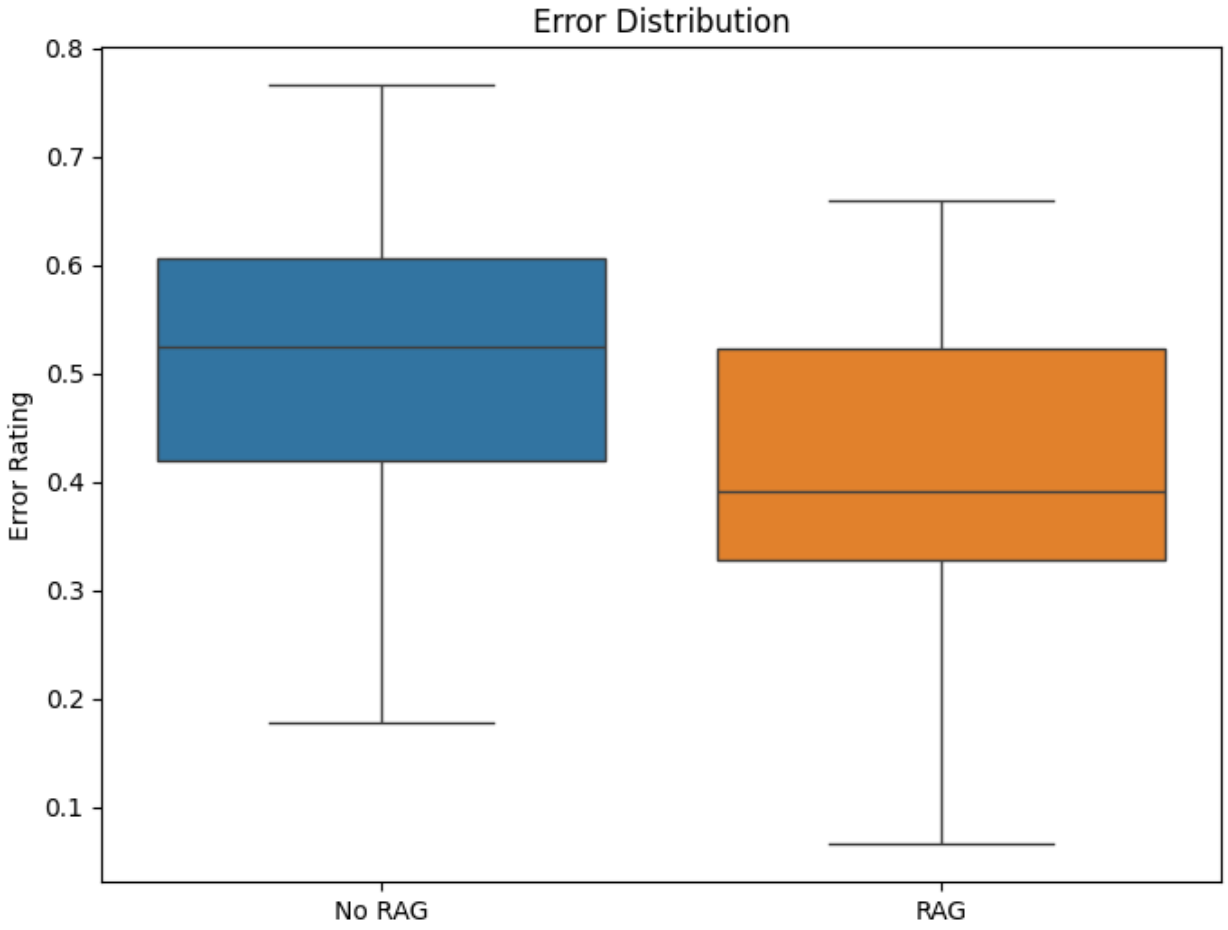
https://github.com/SebastianRosenquist/QRM_LLM_Scraper

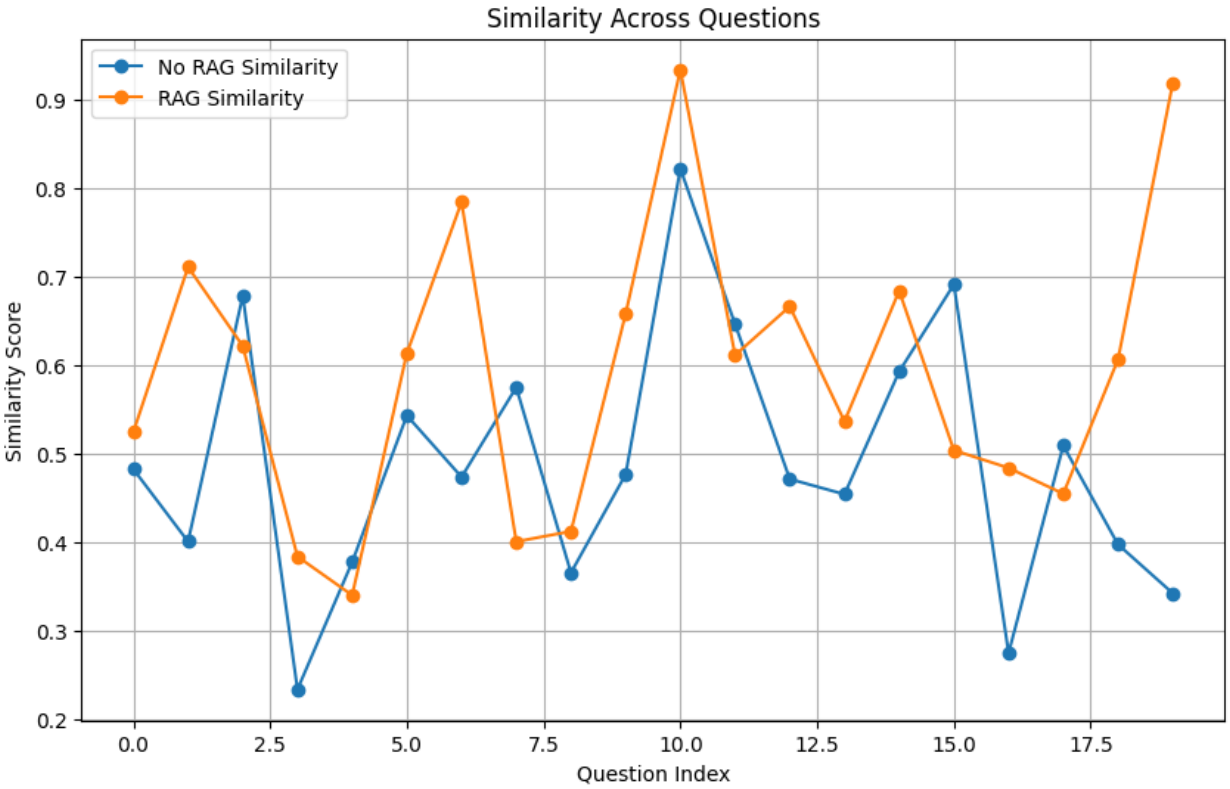
8.2. MORE FINDINGS




Average Error
Error Rating
0.0
0.1
0.2
0.3
0.4
0.5
No RAG Error
RAG Error


Similarity Distribution
Similarity Score
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
No RAG
RAG


Error Distribution
Error Rating
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
No RAG
RAG


Similarity Across Questions
No RAG Similarity
RAG Similarity
Similarity Score
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Question Index