Northeastern University

# Can AI Be Trusted? A Study on Reliability in Retrieval-Augmented Generation Responses

Author: **Mulei Ni**
ID: 002298788
E-mail: *ni.mul@northeastern.edu*

Author: **Hejia Li**
ID: 002640164
E-mail: *li.hej@northeastern.edu*

Author: **Zheng Gu**
ID: 002647508
E-mail: *gu.zhen@northeastern.edu*

December 3, 2024

# Contents

1 Introduction 2
1.1 The RBCM AI Avatar Project . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2
1.2 Importance . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2
1.3 Research Gaps . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2

2 Proposed Solutions 3
2.1 Related Work . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3
2.2 Solutions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4
2.2.1 Human Evaluation . . . . . . . . . . . . . . . . . . . . . . . . . . 4
2.2.2 LLM-as-a-Judge . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

3 Methodology 7
3.1 Literature Review . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7
3.1.1 RAG Evaluation Frameworks . . . . . . . . . . . . . . . . . . . . . 7
3.1.2 Sustainable RAG . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8
3.2 Plan for Experiments . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8
3.2.1 Naive RAG Model . . . . . . . . . . . . . . . . . . . . . . . . . . . 8
3.2.2 Dataset . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9
3.2.3 Experiments . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10
3.2.4 Evaluation Framework and Metrics . . . . . . . . . . . . . . . . . . 11
3.2.5 Green Software Evaluation . . . . . . . . . . . . . . . . . . . . . . 11

4 Discussion: Results Analysis 12
4.1 Feasibility of LLM-as-a-Judge . . . . . . . . . . . . . . . . . . . . . . . . . 12
4.2 Raw Models and RAG Evaluation Results . . . . . . . . . . . . . . . . . . . 12
4.3 Carbon Emissions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13

5 Future Work 13

# 1 Introduction

The rise of ChatGPT and other Artificial Intelligence (AI) technologies has pushed Large Language Models (LLMs) to the forefront of the tech industry and sparked significant advancements and discussions. However, as these models evolve, there remains a critical need for rigorous evaluation to ensure their reliability and effectiveness. This research, conducted as part of the Royal British Columbia Museum (RBCM) AI Avatar Project, explores comprehensive methodologies for evaluating AI-generated responses. Furthermore, given the substantial computational demands of training and deploying LLMs, this study integrates energy consumption metrics into the evaluation process, emphasizing the importance of sustainable AI practices and providing evidence for environmentally conscious innovation.

## 1.1 The RBCM AI Avatar Project

The RBCM is leveraging AI avatars to transform visitor experiences by delivering personalized, interactive guidance. Utilizing Retrieval-Augmented Generation (RAG), these avatars tap into the museum's vast collection to provide contextually relevant responses to visitor inquiries. While RAG holds significant potential, evaluating the quality, accuracy, and contextual appropriateness of the generated responses presents a considerable challenge due to the inherent complexities of RAG systems.

## 1.2 Importance

The deployment of AI avatars offers museums a powerful tool to enhance visitor engagement, share cultural preservation, and extend the educational impact of exhibits. Acting as virtual docents, AI avatars can offer personalized tours, deliver detailed explanations, and make museum content more accessible to diverse audiences. These digital guides also play a critical role in advancing museums' educational missions, enabling tailored learning experiences that deepen understanding of art, history, and culture. Additionally, AI avatars contribute to the creation and enrichment of digital archives, ensuring that valuable knowledge is preserved and shared with future generations.

Despite their potential, AI avatars present significant challenges, particularly in ensuring the accuracy and reliability of the information they provide. In a museum context, any misleading or incorrect information—referred to as "hallucinations"—can result in significant misinterpretations of cultural and historical context, negatively impacting the visitor experience. This issue underscores the importance of developing effective evaluation methods, forming the foundation for this research.

## 1.3 Research Gaps

LLMs face persistent challenges, such as hallucinations and inaccuracies [14], [15]. These limitations are particularly critical in contexts like museums, where accuracy and reliability are crucial. Despite the growing adoption of RAG systems, systematic approaches for validating AI-generated responses remain underdeveloped. Several key research gaps emerge in the evaluation of RAG responses:

- **Limited Research on Holistic RAG Evaluations**: Existing studies largely focus on individual components of RAG systems, such as retrieval or generation, without addressing the overall performance of the system [13], [5]. This approach prevents a comprehensive understanding of the overall effectiveness of RAG. We need a holistic evaluation framework to assess the quality, accuracy, relevance, and contextual appropriateness of the entire RAG pipeline.
- **Lack of Standardized Benchmark**: While various metrics and frameworks exist for evaluating RAG systems, there is no universally accepted benchmark or unified metric set. Differences in evaluation criteria, such as accuracy, precision, and recall, complicate cross-comparisons and make it difficult to establish consistent standards for assessing model performance [16], [17].
- **Lack of Museum-Specific Datasets**: There is a distinct lack of museum domain-specific datasets, which prevents the ability to effectively benchmark RAG responses. Without domain-specific datasets, ensuring that RAG-generated responses are both factual and contextually relevant remains a challenge.
- **Expensive Human Evaluation**: Human evaluation, though widely used, is resource-intensive and relies on domain experts manually scoring RAG-generated responses [22]. This method is not scalable, especially for niche domains like museums.

To address these gaps, this research will examine the existing RAG evaluation pipeline and investigate the emerging approach of utilizing the "LLM-as-a-Judge" highlighted in recent research. Experiments will be conducted to assess the effectiveness of this approach in evaluating RAG systems. Additionally, given the growing concern over the substantial energy demands of LLM training and deployment, this study will assess the carbon emissions associated with the RAG process to promote sustainable AI practices.

# 2 Proposed Solutions

## 2.1 Related Work

The performance of Retrieval-Augmented Generation (RAG) models is influenced by several factors, with chunk size being one of the most critical. Antonio Jimeno-Yepes et al. conducted an in-depth analysis of chunking strategies within the RAG process, comparing them to a baseline case [2]. They evaluated accuracy using ROUGE and BLEU scores, offering valuable insights into how different chunking approaches can enhance retrieval performance. This analysis is particularly relevant to our project, as it underscores the importance of effective chunking in improving RAG outcomes.

Another significant factor is the choice of embedding models. Caspari et al. highlighted the critical role of embedding model selection, evaluating performance through two methods: assessing embedding alignment using Centered Kernel Alignment and comparing retrieval results based on Jaccard and rank similarity [3]. Their research expanded the evaluation paradigm beyond standard benchmarks, providing a more nuanced understanding of embedding model efficacy.

To address issues like hallucinations and inaccuracies in LLM outputs, RAG systems retrieve information from external knowledge sources. Jiang et al. observed that most current RAG systems retrieve information based solely on input, which limits their effectiveness in generating longer texts [4]. They proposed adjustments to the generation process to improve RAG performance, particularly for extended responses.

The challenge of quantitatively evaluating RAG models has also been a focus of recent research. Dongyu Ru et al. assessed eight RAG systems and demonstrated that RAGChecker exhibits strong correlations with human judgments, offering a reliable framework for RAG evaluation [5]. This approach addresses the inherent difficulties in benchmarking RAG systems against human evaluation standards.

Further optimizations to the RAG workflow have also been explored. Shi et al. systematically investigated best practices across various components of the RAG process, including query classification, retrieval, reranking, and summarization [6]. Their findings revealed that refining these individual modules significantly reduces hallucinations and improves the relevance of generated content. This comprehensive approach demonstrates the potential for modular optimization to enhance the overall performance of RAG systems.

Finally, Wang et al. introduced ERAGent, a novel framework designed to enhance RAG systems by incorporating an Enhanced Question Rewriter and a Personalized LLM Reader [7]. These components refine input queries and integrate user preferences, leading to improved accuracy and personalization in generated responses. Additionally, the Experiential Learner module expands the system's knowledge boundaries by learning from historical dialogues, enhancing both efficiency and adaptability to users' evolving interests.

## 2.2 Solutions

The studies discussed above highlight several factors influencing the performance of RAG-based AI-generated content, including chunk size, embedding model selection, and generation process adjustments. While many novel evaluation frameworks have emerged, direct comparisons between human and automated evaluation approaches remain limited due to the evolving nature of the field. To address this gap, this study evaluates two key approaches to assessing RAG performance: human evaluation, traditionally considered the gold standard, and LLM-based evaluation, an emerging and scalable alternative.

### 2.2.1 Human Evaluation

Human evaluation is widely regarded as the gold standard in assessing the performance of language models, particularly in tasks requiring contextual understanding and nuanced judgment. Human evaluators can interpret complex scenarios and account for subtle intricacies, providing invaluable insights into a model's effectiveness [22]. However, this approach has notable drawbacks.

First, human evaluation is inherently time-consuming and expensive. It requires the involvement of domain experts who can assess the LLM's responses accurately, especially for tasks in specialized fields. Ensuring consistency across human evaluators is another challenge, as subjective judgment can lead to variability in scoring standards. To mitigate this, evaluators must align with strict evaluation guidelines, which increases the complexity and expense of the process.

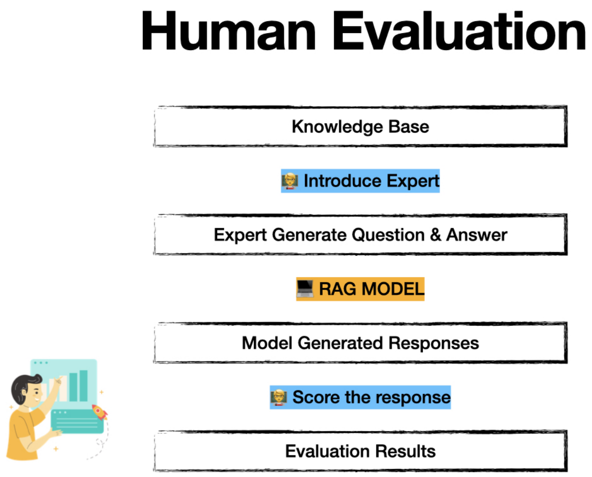

Figure 1: Human Evaluation Pipeline

Second, scalability poses a significant challenge for human evaluation. Real-world applications of LLMs can produce an immense volume of responses, often exceeding 100,000 outputs per month. At an estimated 60 seconds per evaluation, this translates to over 6 million seconds—or nearly 70 days—of continuous work, an unfeasible demand for human evaluators. Such a scale makes relying solely on human judgment impractical in large-scale scenarios [23].

While humans may excel at understanding content within specific domains, their involvement in large-scale evaluations is neither scalable nor cost-effective. As a result, alternative or complementary evaluation methods, such as automated metrics or hybrid systems combining human and machine evaluation, are essential to address the demands of real-world applications. [9].

### 2.2.2 LLM-as-a-Judge

Using large language models (LLMs) as judges provides an innovative and efficient alternative to human evaluation. LLMs can evaluate responses based on predefined, user-tailored criteria, addressing many challenges inherent in human evaluation, such as time and cost constraints. However, this approach raises concerns about potential biases, as LLMs function as both the "player" (response generator) and the "referee" (evaluator). Despite these concerns, research has demonstrated promising results. For instance, studies show that GPT-4 achieves 85% agreement with human evaluations, even exceeding the agreement level among human experts. This underscores the superior consistency of LLMs compared to human evaluators, making them a scalable and reliable alternative [20].

#### The Three-Stage Process

The process of using LLMs as judges includes three distinct stages[24]:

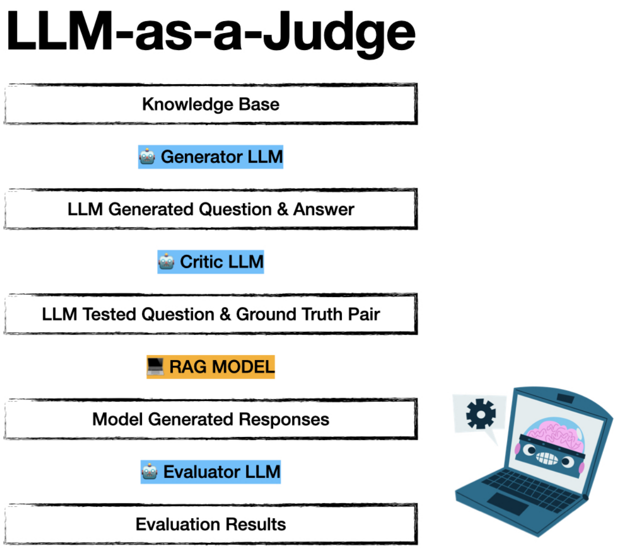
LLM-as-a-Judge
Knowledge Base
Generator LLM
LLM Generated Question & Answer
Critic LLM
LLM Tested Question & Ground Truth Pair
RAG MODEL
Model Generated Responses
Evaluator LLM
Evaluation Results

Figure 2: LLM-as-a-Judge Evaluation Pipeline

1. **Generation:** The first LLM, known as the *generation LLM*, generates question pairs that include queries and corresponding ground truth answers. These pairs form the foundation for subsequent evaluation.
2. **Critique:** The second LLM, referred to as the *critique LLM*, reviews the generated question pairs to filter out irrational or irrelevant ones. This step ensures a high-quality test dataset, enhancing the validity of the evaluation.
3. **Evaluation:** The third LLM, the *evaluation LLM*, is trained specifically to assess the performance of an LLM application. For example, in a Retrieval-Augmented Generation (RAG) system, the test dataset is fed into the RAG system, producing a response. The response, query, and ground truth are then bundled into a package alongside predefined evaluation prompts and a standardized rubric. This package is sent to the evaluation LLM, which generates a final score based on the given criteria.

This multi-step process creates a robust and adaptable pipeline for evaluating LLM performance. By aggregating scores, organizations can benchmark LLM applications, conduct comprehensive evaluations, and perform regression testing. A key advantage of this approach is its flexibility, allowing modification of evaluation prompts and rubrics to enable nuanced, multidimensional assessments tailored to specific applications, such as the RBCM AI Avatar project.

Additionally, incorporating three specialized LLMs into the pipeline addresses concerns like hallucinations and biases that can arise when using LLM-as-a-judge. This systematic approach ensures reliability and scalability, offering a powerful solution for evaluating LLMs in practical, real-world scenarios.

# 3 Methodology

## 3.1 Literature Review

### 3.1.1 RAG Evaluation Frameworks

Several frameworks are available to evaluate the performance of RAG models. Shahul Es et al. demonstrated the effectiveness of **RAGAS**, showing that its evaluations closely align with human predictions, making it a reliable tool and a strong reference for our study [9]. These findings offer valuable insights for our project, which led us to select RAGAS as the primary assessment framework due to its ease of use and reliability.

Subash et al. conducted a similar study in which they developed an LLM-based chatbot, BARKPLUG V.2, using RAG pipelines for Mississippi State University. The chatbot aims to enhance access to university resources by providing interactive, domain-specific question-answering capabilities. The model was evaluated by RAGAS in terms of context precision, context recall, faithfulness, and relevance. The results highlighted the potential of such educational chatbots in delivering precise and contextually relevant responses. However, limitations such as the absence of Automatic Speech Recognition (ASR) and the risk of hallucinations were noted. The study also emphasized that RAGAS is more suitable for RAG-based systems than traditional metrics like ROUGE and BLEU, as it is specifically designed for these pipelines [11].

In response to the rapid development of evaluation frameworks, Sujoy et al. developed an enhanced version of RAGAS to improve transparency in deriving numerical values for evaluation metrics. The study focused on a telecom domain QA database. The results showed that metrics such as Answer Relevance and Context Relevance, which heavily rely on cosine similarity, have limitations in reliability. In contrast, components like faithfulness and factual correctness were found to align closely with expert evaluations, making them more suitable for assessing RAG models within these pipelines [12].

Apart from evaluation frameworks, the feasibility of using LLMs as judges has started to catch researchers' attention. L. Zheng et al. explored this concept by comparing LLM-based evaluations with human evaluations. The study addressed the limitations of existing benchmarks in effectively capturing human preferences, especially in open-ended tasks. The authors proposed methods to mitigate biases such as position bias and verbosity bias. The results indicated that this approach could be a valid evaluation framework, demonstrating a high level of agreement (over 80%) with human experts' judgments [16].

Additionally, P. Qian et al. introduced EvaluLLM, an evaluation tool that leverages LLMs as customizable judges. The study aimed to assess the necessity of human involvement when developing effective evaluation criteria for LLM-based assessments. Experts from eight different domains were interviewed to identify challenges and user needs within such an evaluation workflow. The results included design suggestions for building optimal evaluation tools, emphasizing the importance of balancing cost-efficiency with human oversight to ensure reliable evaluations, particularly in creative or domain-specific tasks [17].

### 3.1.2 Sustainable RAG

In a recent study, Jinbo Wen et al. proposed a carbon emission optimization framework for mobile AI-Generated Content (AIGC) using RAG. The researchers employed CodeCarbon to demonstrate that no excessive carbon emissions occurred during the training process. Optimal results were identified using Generative Diffusion Models (GDMs) [1]. While the study provides valuable insights into the RAG process and offers source code for energy consumption, its limitations lie in its focus on locally trained models and the absence of carbon emission evaluations when using APIs.

Furthermore, Li et al. proposed the SELF-ROUTE method, which dynamically routes queries between RAG and long-context LLMs based on the model's self-assessment [8]. This approach adjusts the generation process by leveraging the strengths of both systems, reducing computational costs while maintaining high performance. Their study underscores the importance of strategic adjustments in the generation workflow to balance efficiency and effectiveness.

Anu et al. discussed critical factors in calculating the energy and carbon emissions associated with using large language models. The study provided detailed calculations to estimate energy consumption per token, inspiring us to consider the calculation methodology and the environmental benefits and cost savings that can arise from using large language models [18].

## 3.2 Plan for Experiments

The RAG evaluation pipeline employs a structured approach to test the system using a knowledge base. Initially, the knowledge base is processed by a basic (naive) model, which generates vectorized representations of its content. These vectors are used to automatically create questions along with corresponding ground truth answers. The generated questions serve as queries, which are then input into the model to produce responses.

This process results in five key components: vectorized documents (linked to their sources), generated queries (questions), ground truths, model responses, and predefined prompts. These components are collectively evaluated using a large language model (LLM), enabling a multidimensional assessment of attributes like faithfulness, relevance, and accuracy.

### 3.2.1 Naive RAG Model

To validate the proposed evaluation methodology, this research employs a naive RAG architecture implemented using LlamaIndex. The choice of a naive setup allows for precise control over key parameters and seamless integration with existing frameworks, facilitating rigorous testing while ensuring the defensibility of the evaluation framework. The RAG approach combines retrieval and generation processes to enable LLMs to incorporate external information into their responses. This architecture is composed of two main components: a retrieval process that fetches relevant data from external sources and a generation process that formulates responses based on the retrieved data. Prior research has demonstrated that even minor changes in these components—such as the selection of the knowledge base or the embedding model—can significantly influence the quality of evaluation outcomes [2], [3], [4].

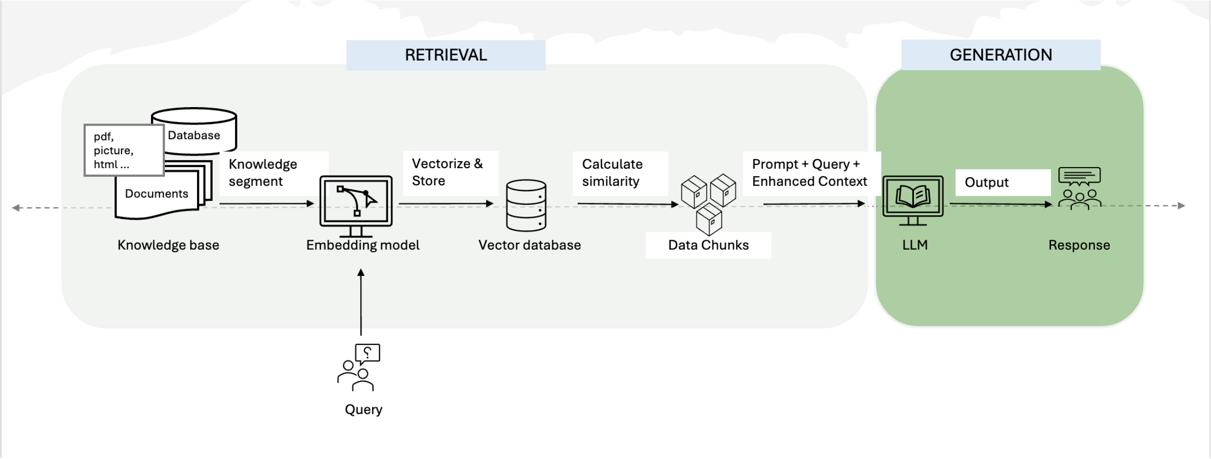

Figure 3: Naive RAG Model

By starting with this naive model, we establish a baseline that isolates the effects of the evaluation framework. This baseline will be valuable for future iterations where more sophisticated variations, such as adjustments to the knowledge base or embedding model, can be introduced to further refine the system and evaluation process. This approach not only ensures the rigor of the testing phase but also sets the stage for iterative improvements based on a solid, controlled foundation.

### 3.2.2 Dataset

A niche dataset, or knowledge base in this context, should ideally have the following key characteristics to efficiently support the LLM-as-a-judge process [25]:

- **Domain-Specific Content**: The knowledge base should focus on specific niches such as finance, law, healthcare, etc., similar to how experts specialize in particular fields. This ensures that the knowledge base is relevant and effective for evaluating domain-specific LLMs.
- **Exclusive Data Sources**: It should contain unique content that is not available in widely used datasets. This reduces redundancy and enhances the model's ability to distinguish itself and perform tasks with higher specificity and depth.
- **Structured as Question & Answer Pairs**: To streamline the model training and evaluation process, the dataset should be organized in the form of clear question-and-answer pairs. This makes it easier to process the data, train the models, and apply standard evaluation methodologies.
- **Accurate and Validated Ground Truth Data**: Each response in the dataset must be fact-checked by domain experts to ensure that the information is both accurate and reliable. This step is crucial for maintaining the credibility of the LLM's evaluations.
- **Traceable Source Passages**: Every question-and-answer pair should be linked to its original source through relevant passage identifiers (e.g.,`relevant_passage_ids`. This ensures transparency and provides a means for validation, allowing evaluators to trace the source material behind each response.

Considering the above requirements, we selected the HotpotQA dataset from Huggingface as the appropriate knowledge base for our experiment. HotpotQA contains 113,000 Wikipedia-based question-answer pairs, with each question-and-answer pair supported by sources and ground truth data [21].

### 3.2.3 Experiments

Upon selection of the knowledge base, we designed two experiments to evaluate the feasibility of using an LLM as a judge:

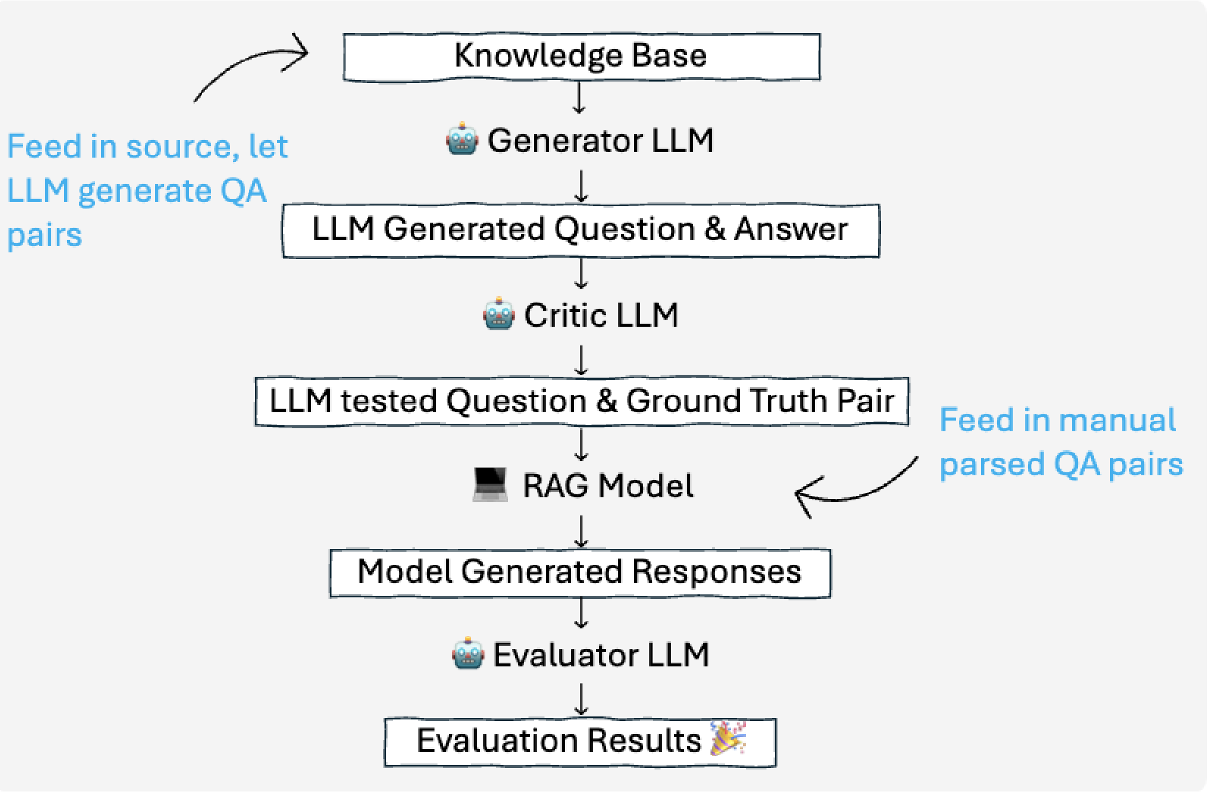
Knowledge Base
Generator LLM
LLM Generated Question & Answer
Critic LLM
LLM tested Question & Ground Truth Pair
RAG Model
Model Generated Responses
Evaluator LLM
Evaluation Results
Feed in source, let LLM generate QA pairs
Feed in manual parsed QA pairs

Figure 4: Control Group Comparison Groups

1. **Control Group Comparison:** Two control groups were tested using a naive RAG model to assess the impact of LLM-based evaluation.
   - LLM-Based Evaluation: The knowledge base was processed by the naive RAG model, and the LLM directly generated evaluation results. This approach involved the generation of question-and-answer pairs by the LLM, which were then used to evaluate the model's performance based on predefined criteria.
   - Pre-Structured Question-and-Answer Evaluation: In this setup, the LLM's question-and-answer generation step was bypassed. Instead, pre-structured question-and-answer pairs were supplied to the model for evaluation.
2. **Model Comparison:** Two different naive models were used as contrast groups within the same evaluation framework. By comparing the results, we aimed to analyze the relationship between RAG models and the evaluation framework. This

comparison helps us understand how variations in the underlying model influence the overall evaluation outcomes and the efficacy of the LLM-as-a-judge approach.

### 3.2.4 Evaluation Framework and Metrics

Based on an extensive literature review, this research selected RAGAS as the evaluation framework due to its ease of use and reliability [9]. RAGAS provides a comprehensive suite of metrics for assessing various dimensions of RAG performance without relying on ground truth human annotations. This framework facilitates faster evaluation cycles for RAG architectures, a critical advantage given the rapid adoption of LLMs.

To comprehensively evaluate the performance of RAG, we employ four key metrics: **Context Precision**, **Context Recall**, **Response Relevancy**, and **Faithfulness**. Each of these metrics assesses distinct aspects of the system's performance, ensuring a holistic evaluation [26].

- **Context Precision**: Evaluates the proportion of relevant information retrieved by the system, focusing on the accuracy of the top-ranked results. It highlights how well the retrieved contexts align with the query.
- **Context Recall**: Measures the system's ability to retrieve all relevant information, emphasizing the completeness of the retrieval process and ensuring no critical information is missed.
- **Response Relevancy**: Assesses how directly and appropriately the generated response addresses the user query, penalizing incomplete or redundant answers to ensure relevance.
- **Faithfulness**: Evaluates the factual consistency of the generated response by verifying whether the claims made in the response can be inferred from the retrieved context, minimizing the risk of hallucinations.

### 3.2.5 Green Software Evaluation

We aim to evaluate the environmental impact (or "greenness") of our naive model in the latter part of the study. We identified the number of tokens used as one of the most important parameters for estimating total carbon emissions, as it is also linked to operational costs [18]. According to ScaleDown, the estimated carbon emission for training GPT-4 is approximately 0.3 grams of CO2 per 1,000 tokens used [19]. This parameter will serve as a baseline for evaluating the model's energy consumption.

To compute the total carbon emissions for our system, we will track the total number of tokens generated during both the training and inference phases. This involves logging the tokens used by the model, multiplying this count by the CO2 emission rate, and then comparing the results with other models or configurations that utilize energy-efficient strategies. Such comparisons will allow us to assess the trade-offs between performance and sustainability.

Additionally, we will identify credible sources to validate and support our approach, adjusting the carbon emission per token as necessary to ensure accurate estimates of the environmental impact. By incorporating these calculations into the evaluation process, we aim to present a comprehensive picture of the energy efficiency of our system, providing insights into how computational efficiency can be balanced with the growing need for sustainable AI practices.

# 4 Discussion: Results Analysis

To validate whether LLM-generated question-ground truth pairs are as effective as human-created ones, we employed two contrasting groups, as detailed in Section 3.2.3. In the first group, the knowledge base was directly integrated into the LLM as part of the evaluation pipeline. In the second group, the data was manually parsed to ensure consistency in the question-ground truth format across both groups. Using the knowledge base described in Section 3.2.2, we aimed to demonstrate the feasibility of utilizing LLM-as-a-Judge within the RAG evaluation pipeline.

## 4.1 Feasibility of LLM-as-a-Judge

The results for the two contrasting groups are summarized in Table 1:

Table 1: Evaluation Results Comparisons between LLM generated dataset and manual parsed datasets

| **Metric** | **Faithfulness** | **Answer Relevancy** | **Context Precision** | **Context Recall** |
|---|---|---|---|---|
| LLM generated QA | 0.794 | 0.972 | 0.684 | 0.870 |
| Manual parsed QA | 0.711 | 0.965 | 1.000 | 0.933 |

The results indicate that LLM-generated QA pairs perform comparably to manually parsed QA pairs across most evaluation metrics when assessed using the RAGAS framework. Metrics such as faithfulness, answer relevancy, and context recall demonstrate similar levels of performance between the two groups.

However, the LLM approach shows a noticeable gap in context precision, which measures whether all relevant ground-truth items are ranked higher in the context. This disparity likely arises from differences in how the contexts are constructed and interpreted by the LLM versus manual parsing. Addressing these discrepancies represents a valuable opportunity for future improvement.

Despite the observed difference in context precision, the comparable performance on other metrics suggests that LLM-generated QA pairs can effectively replace manually created pairs. This substitution has the potential to significantly reduce the time and effort required for human evaluation, validating the feasibility of LLM-as-a-Judge.

## 4.2 Raw Models and RAG Evaluation Results

To explore the relationship between raw models and evaluation results, we evaluated two baseline RAG models: GPT-4o and GPT-3.5 Turbo. Surprisingly, when applied to the larger knowledge base detailed in Section 3.2.2, the more advanced model, GPT-4o, exhibited lower average evaluation scores across several metrics compared to GPT-3.5 Turbo.

Table 2: Mean Relevancy Comparison Between GPT-4o and GPT-3.5-Turbo (Huggingface Dataset)

| **Metric** | **GPT-4o** | **GPT-3.5-Turbo** |
|---|---|---|
| Faithfulness | 0.789 | 0.7875 |
| Answer Relevancy | 0.826 | 0.9185 |
| Context Precision | 0.9 | 0.9 |
| Context Recall | 0.9 | 0.96 |

The comparison highlights that GPT-3.5 Turbo achieved higher scores in metrics such as answer relevancy and context recall, while both models performed equally well in context precision. This finding suggests that the more advanced architecture of GPT-4o might not always translate to better performance in RAG evaluation, particularly when dealing with larger and more complex knowledge bases. These results show the importance of model selection in RAG evaluation pipelines and highlight potential trade-offs between model complexity and performance on specific metrics. Further research could be conducted to understand the underlying factors contributing to these differences.

### 4.3 Carbon Emissions

To estimate carbon emissions, we applied specific equations to calculate the amount of CO2 generated per token. By multiplying this value by the total number of tokens, we derived an estimate of the total carbon emissions produced by each model. Interestingly, our findings indicate that GPT-4o generates slightly fewer carbon emissions compared to GPT-3.5 Turbo, primarily due to its lower token usage. However, it is important to note that carbon emissions per token can vary between models, which may introduce some uncertainty into the emissions estimates.

Table 3: Model Token Count and Carbon Emission Comparison

| **Model** | **Tokens** | **Carbon Emission** |
|---|---|---|
| GPT-4.0 | 231,778 | 69.5334 grams |
| GPT-3.5 | 234,780 | 70.4340 grams |

## 5 Future Work

RAG evaluation is a complex process. While this research has demonstrated the feasibility of using the "LLM-as-a-Judge", several limitations and open questions remain as discussed in Section 4. These present promising directions for future work:

- **Guiding RBCM in Selecting RAG Evaluation Methodologies**: This research concludes that the "LLM-as-a-judge" approach is a feasible alternative to human evaluation. However, for museum-specific contexts, various evaluation strategies may be appropriate. Future work should focus on providing RBCM with comprehensive comparisons of these methodologies, enabling the museum to choose the most suitable solution based on additional research findings and available resources.

- **Optimizing RAG Models with Museum-Specific Datasets**: Datasets play a critical role in the performance of RAG models. While this study utilized a general knowledge base for evaluation, museum-specific datasets may yield different outcomes. Future work should prioritize developing and refining a tailored RBCM dataset, which would enable more accurate benchmarking and provide deeper insights into RAG model performance across museum-specific use cases.
- **Expanding Evaluation Criteria**: This research primarily focused on four key metrics: context precision, context recall, faithfulness, and answer relevancy. Expanding these criteria to include additional metrics, such as noise sensitivity and context entity recall, will allow for a more comprehensive assessment of RAG models and their robustness in varied scenarios.
- **Improving RAG Model Selection**: The experiments conducted did not identify which RAG model performs best. Future studies should focus on creating intensive, domain-specific test cases and investigating advanced methodologies for selecting the most effective RAG models.
- **Incorporating Sustainability into RAG Processes**: With the growing emphasis on sustainability in AI research, future work should explore environmentally-conscious practices for RAG systems. This includes optimizing energy consumption, reducing carbon emissions, and assessing environmental impacts during RAG processes. Such efforts will help align AI advancements with sustainable development goals.

# References

[1] J. Wen et al., “Generative AI for Low-Carbon Artificial Intelligence of Things with Large Language Models,” Jul. 17, 2024, arXiv: arXiv:2404.18077. doi: 10.48550/arXiv.2404.18077.

[2] A. J. Yepes, Y. You, J. Milczek, S. Laverde, and R. Li, “Financial Report Chunking for Effective Retrieval Augmented Generation,” arXiv.org. Accessed: Oct. 02, 2024. [Online]. Available: https://arxiv.org/abs/2402.05131v3

[3] L. Caspari, K. G. Dastidar, S. Zerhoudi, J. Mitrovic, and M. Granitzer, “Beyond Benchmarks: Evaluating Embedding Model Similarity for Retrieval Augmented Generation Systems,” arXiv.org. Accessed: Oct. 02, 2024. [Online]. Available: https://arxiv.org/abs/2407.08275v1

[4] Z. Jiang et al., “Active Retrieval Augmented Generation,” Oct. 21, 2023, arXiv: arXiv:2305.06983. doi: 10.48550/arXiv.2305.06983.

[5] D. Ru et al., “RAGChecker: A Fine-grained Framework for Diagnosing Retrieval-Augmented Generation,” Aug. 16, 2024, arXiv: arXiv:2408.08067. doi: 10.48550/arXiv.2408.08067.

[6] Y. Shi, X. Zi, Z. Shi, H. Zhang, Q. Wu, and M. Xu, "ERAGent: Enhancing Retrieval-Augmented Language Models with Improved Accuracy, Efficiency, and Personalization," May 06, 2024, arXiv: arXiv:2405.06683. doi: 10.48550/arXiv.2405.06683.

[7] X. Wang et al., "Searching for Best Practices in Retrieval-Augmented Generation," Jul. 01, 2024, arXiv: arXiv:2407.01219. doi: 10.48550/arXiv.2407.01219.

[8] Z. Li, C. Li, M. Zhang, Q. Mei, and M. Bendersky, "Retrieval Augmented Generation or Long-Context LLMs? A Comprehensive Study and Hybrid Approach," Jul. 23, 2024, arXiv: arXiv:2407.16833. doi: 10.48550/arXiv.2407.16833.

[9] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated Evaluation of Retrieval Augmented Generation," Sep. 26, 2023, arXiv: arXiv:2309.15217. doi: 10.48550/arXiv.2309.15217.

[10] "Evaluation - LlamaIndex." Accessed: Oct. 09, 2024. [Online]. Available: https://docs.llamaindex.ai/en/stable/optimizing/evaluation/evaluation/

[11] S. Neupane et al., "From Questions to Insightful Answers: Building an Informed Chatbot for University Resources," May 13, 2024, arXiv: arXiv:2405.08120. doi: 10.48550/arXiv.2405.08120.

[12] S. Roychowdhury, S. Soman, H. G. Ranjani, N. Gunda, V. Chhabra, and S. K. Bala, "Evaluation of RAG Metrics for Question Answering in the Telecom Domain," Jul. 15, 2024, arXiv: arXiv:2407.12873. doi: 10.48550/arXiv.2407.12873.

[13] H. Yu, A. Gan, K. Zhang, S. Tong, Q. Liu, and Z. Liu, "Evaluation of Retrieval-Augmented Generation: A Survey," Jul. 03, 2024, arXiv: arXiv:2405.07437. doi: 10.48550/arXiv.2405.07437.

[14] S. M. T. I. Tonmoy et al., "A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models," Jan. 08, 2024, arXiv: arXiv:2401.01313. doi: 10.48550/arXiv.2401.01313.

[15] X. Hu et al., "RefChecker: Reference-based Fine-grained Hallucination Checker and Benchmark for Large Language Models," May 23, 2024, arXiv: arXiv:2405.14486. doi: 10.48550/arXiv.2405.14486.

[16] L. Zheng et al., "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," Dec. 24, 2023, arXiv: arXiv:2306.05685. doi: 10.48550/arXiv.2306.05685.

[17] Q. Pan et al., "Human-Centered Design Recommendations for LLM-as-a-Judge," Jul. 03, 2024, arXiv: arXiv:2407.03479. doi: 10.48550/arXiv.2407.03479.

[18] Anu, "We can use 'tokens' to track AI's carbon emissions: here's how," Anu's Substack. Accessed: Oct. 19, 2024. [Online]. Available: https://anuragsridharan.substack.com/p/we-can-use-tokens-to-track-ais-carbon

[19] Soham, "The cost of inference: Running the Models," The Cost of Inference: Running the Models - by Soham, Accessed: Oct. 19, 2024. [Online]. Available: https://tinyml.substack.com/p/the-cost-of-inference-running-the

[20] Y. Wang, A. G. Hernandez, R. Kyslyi, and N. Kersting, "Evaluating Quality of Answers for Retrieval-Augmented Generation: A Strong LLM Is

All You Need," arXiv.org. Accessed: Oct. 20, 2024. [Online]. Available: https://arxiv.org/abs/2406.18064v2

[21] "rungalileo/ragbench · Datasets at Hugging Face." Accessed: Nov. 19, 2024. [Online]. Available: https://huggingface.co/datasets/rungalileo/ragbench/viewer/hotpotqa/train

[22] "Best Practices for LLM Evaluation of RAG Applications," Databricks. Accessed: Nov. 25, 2024. [Online]. Available: https://www.databricks.com/blog/LLM-auto-eval-best-practices-RAG

[23] "Leveraging LLM-as-a-Judge for Automated and Scalable Evaluation - Confident AI." Accessed: Nov. 27, 2024. [Online]. Available: https://www.confident-ai.com/blog/why-llm-as-a-judge-is-the-best-llm-evaluation-method

[24] "Generate synthetic data for evaluating RAG systems using Amazon Bedrock | AWS Machine Learning Blog." Accessed: Nov. 27, 2024. [Online]. Available: https://aws.amazon.com/blogs/machine-learning/generate-synthetic-data-for-evaluating-rag-systems-using-amazon-bedrock/

[25] K. Zhu et al., "RAGEval: Scenario Specific RAG Evaluation Dataset Generation Framework," Oct. 17, 2024, arXiv: arXiv:2408.01262. doi: 10.48550/arXiv.2408.01262.

[26] "List of available metrics - Ragas." Accessed: Nov. 27, 2024. [Online]. Available: https://docs.ragas.io/en/latest/concepts/metrics/available_metrics/