RAG-ME - Retrieval Augmented Generation Multi-database Evaluation: Towards a Fairer Assessment

Guilherme G. Zanetti*, Thiago M. Paixão†, Filipe W. Mutz*, Sérgio Mucciaccia*,
Claudine S. Badue*, Alberto F. De Souza*, Thiago Oliveira-Santos*
* Universidade Federal do Espírito Santo (UFES); † Instituto Federal do Espírito Santo (IFES)

***Abstract*—Traditional Retrieval Augmented Generation (RAG) system evaluation often suffers from optimistic bias when using a single text source for both question generation and retrieval context. This paper introduces RAG-ME, a framework whose core contribution lies in decoupling the text source for question generation from the source used for retrieval context. RAG-ME employs one source to generate evaluation questions and a distinct, similar source to provide the context for answer retrieval. System-generated answers are then evaluated with LLM-as-a-Judge. Our experiments, conducted on scientific papers and veterinary pathology textbooks, indicate that RAG-ME's dual-database approach yields a fairer assessment, significantly reducing bias and more effectively differentiating model capabilities compared to conventional single-database methods. Furthermore, RAG-ME's capacity of analyzing retrieval parameters (e.g., chunk size and overlap) highlights its practical value for system optimization, offering a cost-effective and reliable tool for advancing fairer RAG assessment.**

I. Introduction

The rise of Large Language Models (LLMs) has revolutionized the field of natural language processing. These models, capable of generating human-like text, have found utility in numerous applications such as automated customer support, code generation, and question answering (QA) [1]. To enhance base models, Retrieval Augmented Generation (RAG) systems have been proposed to contextually ground the answers, avoid hallucinations, and provide information initially unknown to the LLM [2]. However, a new challenge has emerged: **how to automatically test and measure the performance of these advanced question-answering systems?**

One of the main challenges of RAG evaluation is the inherent open-ended nature of question answering in natural language. Most questions, especially complex ones, can have multiple correct answers written in many ways. This characteristic makes classic text comparison metrics, such as BLEU and ROUGE fall short of capturing the nuanced and context-specific quality of responses generated by LLM [3].

To avoid this issue, multiple evaluation methods started to use LLMs in the process. Prometheus [4] proposes the fine-tuning of small generative models to judge the quality of an answer by direct assessment (score from 1 to 5) or to compare two different answers according to predefined criteria. Another approach initially proposed by RAGAS [5] is to use a state-of-the-art (SOTA) LLM to evaluate the components of a RAG system, such as Context Faithfulness, Answer Relevance and Context Relevance. RAGAS also introduced the idea of using LLMs to identify the independent claims made by a text, which can later be compared to the claims in another text for similarity analysis. This same idea was deeper explored by RAGChecker [6], which used fine-grained claim-level checking to measure the performance of the retriever, generator and overall system based on a ground truth answer.

In addition to the metrics used for QA system evaluation, recent works focus on creating synthetic datasets with minimal human oversight. Kenneweg et al. [7] propose a semi-automatic method to create an evaluation dataset from Wikipedia articles written after the LLM's training period, to decrease the influence of the generative model's prior knowledge. Similarly, Eibich et al. [8] compare different RAG methods using a synthetic open-ended dataset of 107 QA pairs from 13 AI arXiv papers generated with GPT-4. Both use SOTA LLMs to score the system's answer to the dataset questions.

Despite these advancements, RAG evaluation remains an open problem, without consensus on the optimal approach. A particularly critical, yet commonly overlooked, limitation in many existing evaluation dataset generation strategies is their reliance on a single text source for both question generation and subsequent retrieval. We hypothesize that this introduces significant biases that can lead to performance overestimation due to the inherent lexical and contextual overlap between the generated questions and the retrieved chunks. Consequently, such evaluation may not accurately reflect a system's robustness or its ability to generalize across diverse phrasings and real-world information retrieval scenarios. Furthermore, this approach often results in an artificially easier evaluation task, making it difficult to discern true differences between high-performing systems and potentially leading to saturated performance metrics.

To directly address this fundamental issue of single-source evaluation bias, this paper introduces RAG-ME (Retrieval Augmented Generation Multi-database Evaluation). The core contribution of RAG-ME lies in its innovative dual-database methodology, which decouples the text source for question generation from the source used for retrieval context. Our framework employs one information source exclusively for generating evaluation questions and a similar source to provide the contextual knowledge base for the RAG system during

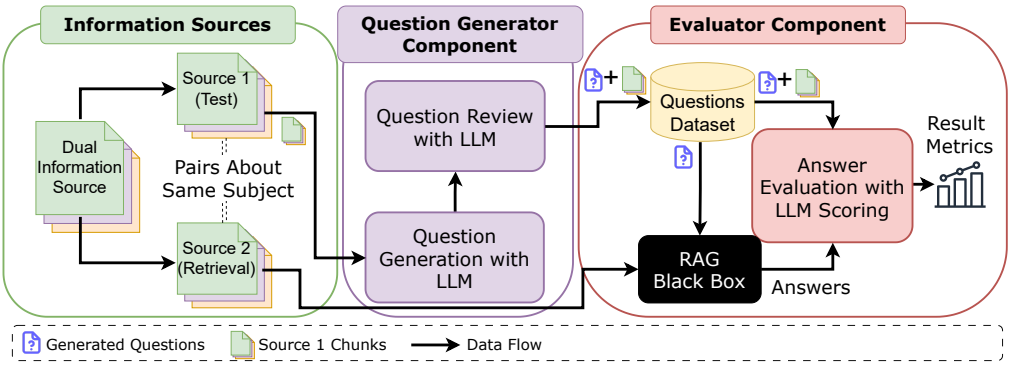

Fig. 1. Flowchart of the proposed framework: The two similar information sources are used separately for the question generation and retrieval augmented answer generation, which go through the evaluation system to generate the final metrics

answer retrieval. Answers produced by the RAG system are then evaluated using a LLM-as-a-Judge [9] scoring mechanism. The proposed method is designed to evaluate black-box RAG systems on any desired knowledge domain, provided two similar knowledge sources are available (e.g., a paper's abstract and its paper body, or two different texts covering the same subject).

In summary, the main contributions of this paper are:

- A novel dual-database evaluation methodology, embodied in the RAG-ME framework, designed to mitigate single-source bias and provide fairer RAG system assessment.
- An empirical analysis demonstrating that this dual-database approach significantly reduces evaluation bias and more effectively differentiates model capabilities compared to conventional single-database methods.
- The code is made available at https://github.com/i2ca/rag-me for reproducibility.

II. The RAG-ME Framework

Figure 1 provides an overview of the RAG-ME framework, illustrating the evaluation process for an arbitrary RAG system. It is assumed that Source 1 and Source 2, two distinct collections of documents, exhibit partial redundancy. This means that the topics explored in Source 1 are also addressed in Source 2, although Source 2 may provide additional content and more comprehensive details on the topics covered by Source 1. The framework includes two main components: automatic **question generator** and **answer evaluator**. The generator component is responsible for producing a dataset with questions based on chunks from Source 1. These questions are then fed into the RAG system with access to Source 2 (operating in a black-box manner), and the generated answers are scored by the evaluator component. The following sections discuss the proposed framework in more detail.

A. Information Sources

The RAG-ME framework relies on two different sources of information. The first source is used for question generation, as seen in Subsection II-B, whereas the second is used by the RAG as the knowledge base to retrieve relevant information. Although information comes from different sources, they should address the same subject and contain similar content.

The proposed dual-source approach aims to reduce the expected biases by using different "learning" and "test" databases (retrieval and question generation, respectively). In this work, as an example, scientific papers were chosen as the main dual information sources, with the abstracts used for question generation and the main body of the paper (paper body, for conciseness) – used as a knowledge base for RAG retrieval. A web scraping script went through the open-access website of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023 (CVPR)[1] and downloaded the PDF files of the published papers. The python library pypdf[2] was used to extract the text from the PDF files, generating a collection of text documents with dual content: the abstract and paper body. Following the convention introduced in Figure 1, the abstracts map to Source 1, whereas the paper bodies map to Source 2.

B. Question Generator Component

The question generator component produces one question for each Source 1 instance (S1-instance). Questions are generated in a multiple-choice format with four answer options, which is required by the quality review system described below. However, it is important to note that the multiple-choice format is used solely for quality assessment during question generation; the actual RAG evaluation (described in Section II-C) uses only the question text itself, with the RAG system generating free-form answers that are then scored by an LLM judge. An example of a generated question is: "*What is a major limitation of existing approaches to visible-infrared recognition?*"

If the S1-instance is longer than 4,000 characters, it is split into chunks of that size to be used as context for the question generation. LLMs are utilized for both generating and reviewing questions.

A preliminary question is generated with Meta's Llama 3 70B for each S1-instance. Although there are lighter versions

[1]https://openaccess.thecvf.com/CVPR2023?day=all
[2]https://github.com/py-pdf/pypdf

of this LLM, the 70 billion parameters version enables the generation of better candidates. The prompt used for the initial generation instructs the model to assume the role of a professor and to generate a multiple choice question with four answer options, answering in a JSON format for easier parsing of the result. The complete prompt is available in the provided code repository.

After the generation step, the question is reviewed in a two-step process. The first step filters out poorly formulated questions that are impossible to answer without additional context. Such questions are incomplete or ambiguous, as they fail to provide sufficient information to be answerable on their own. For example, "*What is a key characteristic of the proposed method for object detection?*" is not answerable because it does not specify which method is being referred to. To identify these problematic questions, the same model used for generation (Llama 3 70B) is employed with a prompt that defines what constitutes a "bad question", presents the question and its answer options, and requests a judgment on whether the question is poorly formulated.

In the second review step, the remaining questions are filtered with the aid of the system proposed by Mucciaccia et al. [10]. This system requires multiple-choice questions with four answer options as input, and it focuses on analyzing question relevance and the quality of answer options. Questions with low relevance or poor-quality answer options are discarded. It is worth mentioning that certain evaluation aspects, such as language, grammar, and format, were disregarded to reduce costs.

C. Evaluator Component

After the generated questions are answered by an arbitrary RAG system, the evaluator component measures its accuracy via LLM-as-a-judge [9]. For each question, the judge LLM (Meta's Llama 3 70B) evaluates the RAG-generated answer by comparing it against the original source text that was used to generate the question. The judge is provided with the question, the original source text, and the RAG's answer, and it follows a chain-of-thought reasoning process to assign an integer score from 1 to 10, reflecting how well the answer addresses the question based on the source information. Answers scoring above 6 are considered correct. The system's overall accuracy is then computed as the percentage of questions that received a passing score. The complete evaluation prompts are available in the provided code repository.

III. Experimental Setup

This section describes experimental setup, including the utilized dataset, the RAG system, and the conducted experiments. All the experiments were run once on a Linux server having one NVIDIA A100 with VRAM of 80GB and 512 GB of RAM. Every open source model was quantized with normalfloat 4bit because of the server's memory constraints.

A. Dataset

To validate the proposed framework, **509 open-access papers** from the 2023 CVPR were downloaded, which are mostly about computer vision techniques and machine learning in general. Using simple regex rules, the abstracts were separated from the paper body, resulting in a collection of 509 abstracts and a collection of 509 paper bodies. If an abstract had more than 4,000 characters, it was divided into 4,000-character chunks, resulting in 512 chunks in total. The chunks from each abstract were then used by a question generator that produced one question. Out of the 512 generated questions, 45 were removed during the review process: 12 were impossible to answer without additional context, 31 had multiple correct options, 1 had no correct options, and 1 was irrelevant. The final dataset contained 463 questions.

A second dataset of 100 questions was generated using two veterinary pathology textbooks from different authors, but covering the same subject. One book was used to generate questions, while the second book was used as the knowledge base for RAG retrieval.

B. Implemented RAG System

The proposed framework was designed to evaluate arbitrary RAG systems; however, this work's experiments use an example RAG system, which is based on a sentence-window approach [8]. In the indexing stage, the documents from the knowledge base are split into text chunks of size c_{size} characters, with an overlap of $c_{\text{overlap}}\%$ between consecutive chunks. Each chunk is then embedded into a vector representation and stored in FAISS vector database [11]. In the retrieval process, the top 3 non-overlapping matches are returned. For each retrieved chunk, the preceding and succeeding chunks are appended, with overlapping portions removed to avoid duplication.

C. Experiments

Four main experiments are conducted to evaluate RAG-ME and test its underlying hypothesis, as described in the following.

1) LLM-as-a-Judge Validation: The purpose of the first experiment is to test whether the LLM-as-a-Judge evaluation component can reliably assess answer correctness. To test this ability and measure its reliability, the experiment compares the LLM's automatic judgments with human expert evaluations based on 40 questions randomly sampled from the CVPR dataset. For each question, one correct and one incorrect answer are generated using a semi-automatic approach. First, the RAG system using *meta-llama/Llama-3.1-70B-Instruct* is prompted to answer the 40 questions automatically. A human expert reviews the responses using two criteria: (1) *factual correctness*: whether the answer is correct according to the source paper, and (2) *completeness*: whether the answer completely answers the question. Based on these criteria, responses are labeled as either *correct* or *incorrect*. Then, for each question, an additional answer is created to ensure that each question has both a correct and an incorrect response: if the RAG-generated answer is correct, the expert creates a wrong but plausible answer; if the RAG-generated answer is incorrect, the human expert manually writes the correct answer. This

yields a collection of 80 question-answer pairs, with each question paired with one correct and one incorrect answer.

The 80-sample collection is evaluated using *LLM-as-a-Judge* with Llama 3 70B as the judge, and compared with human labels using accuracy, F1-score and Pearson Correlation. Additionally, we compare *LLM-as-a-Judge* with the overall score of RAGChecker [6], which could serve as an alternative metric to LLM-as-a-Judge.

2) Single vs. Dual-database Comparison: This experiment tests RAG-ME's main hypothesis: that using a single-database for both question generation and retrieval introduces optimistic bias that overestimates system performance and reduces the ability to differentiate between systems. The purpose is twofold: (1) to test whether single-database evaluation consistently produces inflated accuracy scores, and (2) to test whether the performance gap is not solely due to task difficulty differences (e.g., abstracts being easier than paper body) but reflects a fundamental bias in single-source evaluation. Three different configurations are evaluated: (1) questions generated from abstracts and RAG with access to abstracts (single-database with abstracts), (2) questions generated from abstracts and RAG with access to paper body (dual-database), and (3) questions generated from paper body and RAG with access to paper body (single-database with paper body). The answers are evaluated with LLM-as-a-Judge to generate an accuracy value for each case. To compare the results, two experiments are performed:

1) The questions are answered by a RAG system with weak and strong generation models from three different families: Llama, Gemma, and Qwen. The RAG system has as constants the embedding model intfloat/multilingual-e5-large-instruct, $c_{\text{size}} = 500$ and $c_{\text{overlap}} = 50\%$.
2) The questions are answered by RAG with four different embedding models: nvidia/NV-Embed-v2, Alibaba-NLP/gte-Qwen2-7B-instruct, intfloat/multilingual-e5-large-instruct, and sentence-transformers/all-MiniLM-L6-v2. The RAG systems use *meta-llama/Llama-3.1-70B-Instruct* as generation model with the same parameters as before. This experiment compares only the dual-database and single-database (abstracts) configurations.

If the hypothesis is correct, it is expected to see higher scores for both single-database configurations compared to the dual-database, and to have more significant difference between the better and worse models in the dual database, both in generation and retrieval.

3) Cross-Domain Generalizability: The purpose of this experiment is to test whether RAG-ME's dual-database methodology generalizes beyond the CVPR computer vision domain. To achieve this, this experiment uses two veterinary pathology textbooks as the dual-database. The same set of generative models used in the CVPR experiments (Llama, Gemma, and Qwen families) are evaluated using both single-database (using a single book) and dual-database configurations. This experiment tests whether the framework can be applied to different domains and text types, not just scientific papers, and whether the single-database bias persists across domains.

TABLE I
AUTOMATIC METRICS EVALUATION WITH HUMAN AS GROUND TRUTH.

Metric	Accuracy	F1-Score	Pearson
LLM-as-a-Judge	85.00%	84.96%	0.703
RAGChecker	73.75%	72.79%	0.512

4) RAG Parameters Analysis: This experiment tests whether RAG-ME can serve as a practical tool for RAG system development and optimization. This case study shows how the framework can be used to make informed decisions about system design choices. The RAG system with generation model *meta-llama/Meta-Llama-3-8B-Instruct* and embedding model *intfloat/multilingual-e5-large-instruct* is run multiple times with varying parameters of c_{size} and c_{overlap}. The values of c_{size} used are 100, 500, 1000 and 5000 characters, and c_{overlap} is chosen among 0, 25, 50, and 75 p.p. of overlap.

IV. EXPERIMENTAL RESULTS

A. LLM-as-a-Judge Validation Results

Table I shows the results for the *LLM-as-a-Judge* validation. The table measures the metrics' accordance with the human expert evaluation. It can be seen that the chosen scoring method for RAG-ME (*LLM-as-a-Judge*) has a high overlap with human evaluation, achieving 85.00% accuracy, 84.96% F1-Score, and a Pearson correlation of 0.703.

The usage of *LLM-as-a-Judge* in RAG-ME results in much less time and cost than manual evaluation, while slightly underestimating the accuracy results of the evaluated systems. Although not perfect, because of the inherent difficulties of the problem, this approach presents a good balance between cost, speed and confidence compared to pure human evaluation and alternative state-of-the-art metrics such as RAGChecker.

B. Single vs. Dual-database Comparison Results

The results of the comparison in Table II show that there is a significant difference between using single-database or dual-database approaches. For the interpretation of these results, we adopt the premise that the first model in each family (Llama-2-7b, gemma-7b, Qwen1.5-7B) is weaker than the corresponding second model (Llama-3.1-70B, gemma-2-27b, Qwen2.5-32B), based on established benchmarks and the LLM Leaderboard [12], which aggregates multiple evaluation benchmarks for generative models. This premise allows us to assess whether the evaluation methods correctly differentiate between model capabilities.

Table II presents the results comparing three evaluation configurations: dual-database (abstracts for questions, paper body for retrieval), single-database with abstracts, and single-database with paper body. The first pattern observed is that both single-database configurations exhibit significantly higher accuracies than the proposed dual-database method, especially for weaker models. This indicates that using a single database

TABLE II
ACCURACY COMPARISON OF THREE EVALUATION CONFIGURATIONS (%).

Model	Dual	Single (abstracts)	Single (body)
Llama-2-7b	52.92	84.88	66.84
Llama-3.1-70B	69.76	95.46	77.55
gemma-7b	28.73	50.11	38.78
gemma-2-27b	66.74	95.90	82.65
Qwen1.5-7B	65.87	93.30	75.00
Qwen2.5-32B	75.59	95.68	83.67

is an easier task when compared to the dual-database, which more accurately simulates a real-world scenario.

To verify that the performance gap is not solely due to the use of abstracts as the single source (which could be considered an easier task), the results of the single-database (paper body) is also presented. The results show that while using the paper body for retrieval is indeed a harder task than using only abstracts (as evidenced by lower scores in the single-database paper body setup compared to single-database abstracts), the single-database paper body results still remain higher than the dual-database results. This suggests that the performance gap we identified is not solely due to the relative simplicity of abstract-based retrieval, but rather reflects a fundamental bias in single-source evaluation methods. This bias can be explained by contextual and lexical similarities between the questions and the chunks they originated from, which facilitates the task of semantic search and answer generation.

Besides the optimistic bias identified in single-database, it can be seen that the dual-database has more significant performance gains across different generative models, with all stronger models outperforming the weaker ones by substantial margins. The same is not true for the conventional single-database methods, which showed smaller differences between weaker and stronger models, resulting in a less informative comparison. In particular, we highlight the result involving Qwen models, where Qwen1.5-7B scored only 2.38 p.p. less than Qwen2.5-32B in the single-database abstracts approach, compared to a 9.72 p.p. difference in the dual-database approach. This Qwen example and the overall results of the newer models in the single-database configurations show signs of a ceiling effect, where performance on a task is so high that it is hard to differentiate between different systems. The dual-database approach, which makes the task more challenging, also helps in this aspect, shown by the higher variation in accuracies among the newer models.

Table III shows the results of the second experiment comparing embedding models across single-database (abstracts) and dual-database approaches. The models are listed in descending order based on their evaluation in MTEB's Retrieval Score [13], a community leaderboard for embedding models similar to the LLM Leaderboard. The accuracy reported is based on the final answer generated using the context from the top 5 retrieved documents, evaluated by LLM-as-a-Judge. It was expected that RAG systems with better embedding models would achieve higher performance, which was partially confirmed in the dual-database results: models with higher MTEB scores generally exhibited higher accuracies in RAG-ME. The exception was gte-Qwen2, which performed slightly worse than multilingual-e5-large-instruct.

TABLE III
ACCURACY COMPARISON OF EMBEDDING MODELS: SINGLE VS. DUAL-DATABASE (%).

Model	Single	Dual	MTEB
NV-Embed-v2	97.19	70.84	72.31
gte-Qwen2-7B-instruct	93.52	66.74	70.24
multilingual-e5-large	95.46	69.76	63.61
all-MiniLM-L6-v2	90.50	62.20	56.09

TABLE IV
ACCURACY COMPARISON ON VETERINARY PATHOLOGY TEXTBOOKS DOMAIN (%).

Model	Single-base	Dual-base
Llama-2-7b	70.71	28.29
Llama-3.1-70B	82.83	43.43
gemma-7b	37.37	10.10
gemma-2-27b	85.86	54.54
Qwen1.5-7B	83.84	39.40
Qwen2.5-32B	91.92	61.62

It can be seen that single-database accuracies were significantly overestimated again, being around 20 to 25 p.p. higher than the dual-database. However, both methods agreed on the same ranking of embedding models, however, it is important to note that the best models achieved accuracies close to 100% on the single-database, which would cause problems in comparison as better models are released.

C. Cross-Domain Generalizability Results

To validate that the RAG-ME framework generalizes beyond the CVPR computer vision domain, we evaluated the same generative models on veterinary pathology textbooks. Table IV presents the results comparing single-database and dual-database configurations in this different domain. The results confirm our original findings: the single-database setup consistently overestimates RAG performance compared to the dual-database approach, indicating that single-database bias persist across different domains. The performance gap between single and dual-database configurations is even more pronounced in this domain, with differences ranging from approximately 30 to 50 p.p.. Additionally, the difference between models of the same family (e.g., Llama-2-7b vs Llama-3.1-70B, Qwen1.5-7B vs Qwen2.5-32B) was also larger in the dual-database scenario than in the single-database, further supporting the claim that dual-database evaluation better differentiates model capabilities.

D. RAG Parameters Analysis Results

As a case study, RAG-ME was used to compare sixteen different parameter combinations of an arbitrary RAG system,

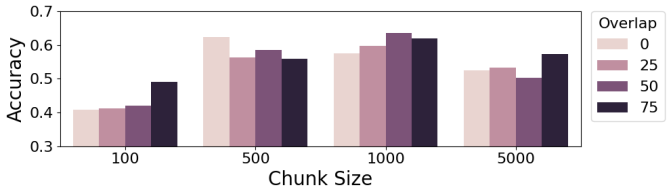

Fig. 2. Bar Plot of accuracies of RAG Systems with different Chunk Sizes and Chunk Overlaps. The x-axis divides the results in blocks by chunk size, and the colors indicate the overlap percentage. Accuracy measured with RAG-ME framework

as shown in Figure 2. The results indicate that a small chunk size (c_{size}) of 100 characters led to significantly worse performance, possibly because such small chunks do not contain enough information for the LLM to answer questions accurately. Conversely, when the chunk size becomes too large, performance also declines. A plausible explanation for this phenomenon is that embedding models struggle to encode the full meaning of excessively large chunks, thereby impairing the retrieval effectiveness of the system.

The analysis of the chunk overlap (c_{overlap}) is not as conclusive. It could be said that bigger chunk overlaps improved the performance of the 100 character chunks, but the other results are inconclusive and seem to indicate that the overlap between chunks does not have a big influence on the performance of a RAG System.

With a similar experiment using RAG-ME, it is possible to choose the best parameters for a specific RAG system in a production environment - in this example, an intermediate chunk size of 500 or 1000 could be chosen. This demonstrates the capability of our framework to help develop and choose new models, RAG techniques and parameters while having more confidence in the evaluation results and a more significant distance between the different systems being compared.

V. Conclusion

This paper introduces RAG-ME, an evaluation framework that addresses bias in RAG system assessment by decoupling the information source for question generation from that used for retrieval context. Multi-domain evaluation indicated that the dual-database methodology mitigates the optimistic bias inherent in single-database evaluation, where lexical and contextual overlap between questions and retrieval sources artificially inflates performance metrics. The dual-database approach not only provides more realistic performance estimates but also better differentiates between model capabilities, reducing ceiling effects that obscure performance differences in single-database settings. The framework's evaluation component, *LLM-as-a-Judge*, achieved 85% agreement with human evaluation, providing a scalable and cost-effective alternative to manual assessment, but it could be replaced by a better SOTA metric, if available in the future.

Aside from the observed benefits, the framework has some limitations that could be addressed in future works. It requires two similar but distinct knowledge sources covering the same subjects, which may not be available for specialized domains or single-source scenarios. The LLM-based scoring, while correlated with human judgment, may introduce biases or miss nuanced aspects of answer quality. Finally, the computational and financial cost of multiple LLM calls for generation and evaluation may limit scalability for large-scale benchmarks.

Despite these limitations, RAG-ME represents a significant step toward fairer and more reliable RAG system evaluation, offering researchers and practitioners a practical tool to assess and optimize retrieval-augmented generation systems with greater confidence in their real-world performance.

Acknowledgments

This study was financed in part by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Brazil) Finance Code 001; Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq); and Fundação de Amparo à Pesquisa e Inovação do Espírito Santo (FAPES, Brazil) - grant 2021-07KJ2 and additional support for event participation (Edital Fapes N° 02/2026).

References

[1] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu, Z. Wu, L. Zhao, D. Zhu, X. Li, N. Qiang, D. Shen, T. Liu, and B. Ge, "Summary of chatgpt-related research and perspective towards the future of large language models," *Meta-Radiology*, vol. 1, no. 2, p. 100017, Sept. 2023.

[2] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Inf. Process. Syst.*, vol. 33, pp. 9459–9474, 2020.

[3] M. Hanna and O. Bojar, "A fine-grained analysis of BERTScore," in *Proc. Sixth Conf. Machine Translation*. Association for Computational Linguistics, Nov. 2021.

[4] S. Kim, J. Suk, S. Longpre, B. Y. Lin, J. Shin, S. Welleck, G. Neubig, M. Lee, K. Lee, and M. Seo, "Prometheus 2: An open source language model specialized in evaluating other language models," in *Proc. 2024 Conf. Empirical Methods Natural Language Processing*. Association for Computational Linguistics, Nov. 2024.

[5] S. Es, J. James, L. Espinosa Anke, and S. Schockaert, "RAGAs: Automated evaluation of retrieval augmented generation," in *Proc. 18th Conf. European Chapter Association Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Mar. 2024.

[6] D. Ru, L. Qiu, X. Hu, T. Zhang, P. Shi, S. Chang, C. Jiayang, C. Wang, S. Sun, H. Li, *et al.*, "Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation," *Advances in Neural Inf. Process. Syst.*, 2024.

[7] T. Kenneweg, P. Kenneweg, and B. Hammer, "Retrieval augmented generation systems: Automatic dataset creation, evaluation and boolean agent setup," 2024.

[8] M. Eibich, S. Nagpal, and A. Fred-Ojala, "ARAGOG: Advanced RAG Output Grading," Apr. 2024.

[9] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging llm-as-a-judge with mt-bench and chatbot arena," in *Proc. 37th Int. Conf. Neural Information Processing Systems*, ser. NIPS '23, 2023.

[10] S. S. Mucciaccia, T. Meireles Paixão, F. Wall Mutz, C. Santos Badue, A. Ferreira de Souza, and T. Oliveira-Santos, "Automatic multiple-choice question generation and evaluation systems based on LLM: A study case with university resolutions," in *Proc. 31st Int. Conf. Computational Linguistics*, Jan. 2025.

[11] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou, "The faiss library," 2024.

[12] E. Beeching, C. Fourrier, N. Habib, S. Han, N. Lambert, N. Rajani, O. Sanseviero, L. Tunstall, and T. Wolf, "Open llm leaderboard," 2023.

[13] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, "Mteb: Massive text embedding benchmark," in *Proc. 17th Conf. European Chapter Association Computational Linguistics*, 2023, pp. 2014–2037.