Evaluating RAG-based QA Systems: A Comparative Analysis of LLM as a Judge, Traditional Metrics, and Human Alignment

Renato Miyaji[1], **Renato Moulin**[1], **Samuel Monção** [1], **Leonardo Machado**[1]

[1]Visagio Group
São Paulo – SP – Brazil

{renato.miyaji, renato.moulin, samuel.moncao, leonardo.machado}@visagio.com

Abstract. *Evaluating RAG based Question Answering systems presents ongoing challenges, as traditional NLP metrics often inadequately capture nuanced answer quality and the reliability of LLM as a Judge paradigms requires further validation. This study comprehensively compares two distinct RAG QA systems on a domain-specific dataset from a consulting company. We investigate the efficacy and human alignment of LLM as a Judge configurations (Fine-Tuning and In-Context Learning) benchmarking them against NLP metrics and human evaluations. Results indicate that BERTScore is more indicative of semantic similarity than lexical-based metrics. For LLM as a Judge evaluations, Prometheus 2 using Pairwise Comparison demonstrated the strongest human alignment.*

1. Introduction

Evaluating and comparing the performance of Large Language Model-based Question Answering (QA) systems remains a challenging task in the literature [Kamalloo et al. 2023]. Although several established metrics such as Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [Lin 2004], Bilingual Evaluation Understudy (BLEU) [Papineni et al. 2002], Metric for Evaluation of Translation with Explicit ORdering (METEOR) [Banerjee and Lavie 2005], and BERTScore [Zhang et al. 2020] are widely used in Natural Language Processing (NLP), their applicability and reliability in specific contexts—such as QA over proprietary or domain-specific documents—are still under investigation [Sai et al. 2022].

Human annotation is often regarded as the gold standard for evaluation, offering the most reliable judgments of answer quality [Sai et al. 2022]. However, this approach faces critical limitations in terms of scalability, cost, and consistency. To address these issues, recent research has increasingly turned to automated evaluation using LLMs themselves—a paradigm commonly referred to as LLM as a Judge [Zheng et al. 2023].

Despite its growing popularity, the use of LLMs for automatic evaluation introduces several open questions [Gu et al. 2024]. These include uncertainties around evaluation formats [Zhu et al. 2025] (Direct Assessment, Pairwise Comparison [Qin et al. 2024]), the choice of scoring criteria (helpfulness, factuality, relevance) [Bai et al. 2023], the underlying methodology (In-Context Learning, Fine-Tuning), and most importantly, the alignment between LLM-based judgments and human evaluations [Gu et al. 2024].

In this context, the present study aims to evaluate and compare two distinct Retrieval-Augmented Generation-based QA systems, using a case study from a consulting company involving real-world documents and user queries. We explore the feasibility and reliability of different LLM as a Judge configurations and assess their agreement with human judgments.

Furthermore, we investigate whether traditional NLP metrics still provide meaningful signals for evaluating QA responses in this setting. By contrasting these metrics with both human annotations and LLM-based evaluations, we aim to provide insights into the strengths and limitations of current evaluation methodologies for LLM-based QA systems.

2. Related Works

2.1. NLP Metrics

Traditional metrics for evaluating NLP tasks have long been established in the literature, particularly for tasks involving text generation and machine translation [Sai et al. 2022]. Among the most widely adopted are BLEU, ROUGE, and METEOR. These metrics rely primarily on surface-level lexical overlaps between generated outputs and reference texts, offering a simple and interpretable way to quantify similarity [Sai et al. 2022].

Bilingual Evaluation Understudy (BLEU) [Papineni et al. 2002] measures the precision of n-grams in the generated text compared to one or more reference texts. Despite its widespread use, BLEU has been criticized for its lack of sensitivity to word order and semantic meaning, often failing to capture paraphrased or semantically equivalent responses [Sai et al. 2022]. Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [Lin 2004], on the other hand, focuses more on recall, computing overlaps of unigrams, bigrams, or longest common subsequences. While more sensitive to content coverage, ROUGE shares similar limitations in failing to account for synonymy or contextual meaning [Schluter 2017].

Metric for Evaluation of Translation with Explicit ORdering (METEOR) [Banerjee and Lavie 2005] was introduced to address some of these shortcomings by incorporating stemming and synonym matching. It aligns words between generated and reference texts and calculates a weighted F-score that balances precision and recall. METEOR tends to correlate better with human judgments than BLEU or ROUGE in certain settings. However, it still depends heavily on near-exact matches and is limited in capturing deeper semantic equivalence [Sai et al. 2022].

To overcome the limitations of these metrics, newer approaches have been developed that leverage embeddings derived from large pretrained Language Models [Sai et al. 2022]. One of the most prominent of these is BERTScore [Zhang et al. 2020], which computes similarity scores between tokens in generated and reference texts using contextualized embeddings from models, such as BERT. By aligning embeddings rather than words, BERTScore captures semantic similarity even when there is no lexical overlap, making it particularly useful in evaluating paraphrased or more abstractive responses.

Despite its advantages, BERTScore is not without limitations. It relies heavily on the quality and domain relevance of the pretrained Language Model used to generate embeddings [Sai et al. 2022]. Moreover, BERTScore may struggle to reflect nuanced

aspects of response quality such as factual consistency, logical coherence, or answer helpfulness—factors that are increasingly important in evaluating QA systems based on Large Language Models. As such, while embedding-based metrics mark a significant advancement over traditional approaches, they still fall short of fully replicating human judgment in complex NLP tasks [Sai et al. 2022].

2.2. LLM as a Judge

An emerging line of research in NLP evaluation involves leveraging LLMs themselves as evaluators—a paradigm commonly referred to as LLM as a Judge [Zheng et al. 2023]. This approach has gained popularity due to its potential to address scalability issues inherent in human annotation [Gu et al. 2024]. By automating the evaluation process, researchers can rapidly assess the quality of generated outputs across large datasets. However, this gain in automation introduces a critical trade-off: ensuring that the evaluations made by LLMs are well-aligned with human judgment remains an ongoing challenge [Yu et al. 2024] [Ho et al. 2025].

To improve this alignment, different training and prompting strategies have been proposed. One common approach is In-Context Learning (ICL), in which the LLM is provided with a few examples of high-quality and low-quality outputs, along with their corresponding scores or labels, within the same prompt [Gu et al. 2024]. This method enables the model to mimic the judging pattern without modifying its parameters. In contrast, Fine-Tuning involves training the LLM on a curated dataset of labeled examples to directly optimize for judgment capabilities [Wang et al. 2023]. While Fine-Tuning can lead to more consistent and robust evaluators, it requires access to large quantities of high-quality annotated data and may be less adaptable across domains.

In addition to training strategies, the evaluation format also plays a key role in the performance of LLM-based judges. Two of the most common methods in the literature are Direct Assessment and Pairwise Comparison [Gu et al. 2024]. In Direct Assessment, the LLM is asked to rate a given response on an absolute scale based on specific criteria, such as relevance, fluency, or factuality [Gu et al. 2024]. In Pairwise Comparison, the model is presented with two candidate responses to the same input and asked to choose which one is better [Zhu et al. 2025]. This relative comparison is often considered more cognitively natural and has shown higher alignment with human preferences in some settings [Kim et al. 2024].

Despite its advantages, Pairwise Comparison introduces its own set of concerns, particularly related to evaluation biases. One well-documented issue is position bias [Ko et al. 2020], where the model systematically favors the response shown in a particular position. Another is format bias [Zhu et al. 2025], in which differences in formatting, length, or structure between responses influence the model's judgment independently of content quality. These biases can distort evaluation outcomes and reduce the reliability of automated assessments.

To mitigate such biases, recent studies have proposed strategies such as randomizing response order, enforcing uniform formatting, or training specialized models to detect and correct these tendencies [Gu et al. 2024]. Nonetheless, ensuring fairness and robustness in LLM-based evaluation remains a complex and unresolved problem. Researchers must carefully design experimental setups and controls to avoid overestimating the effec-

tiveness of LLM judges.

Another important consideration is the definition of evaluation criteria used during the judgment process [Kim et al. 2024]. Depending on the task, criteria may include relevance, factual accuracy, coherence, fluency, and helpfulness, among others [Gu et al. 2024]. Selecting and articulating these criteria clearly in prompts is crucial for guiding the model's judgment and achieving consistent results.

Overall, while the use of LLMs as evaluators holds promise for scalable and cost-effective QA system evaluation, it also raises fundamental questions about reliability [Yu et al. 2024], bias [Zhu et al. 2025], and human alignment [Kim et al. 2024]. The current literature presents a growing body of work seeking to address these issues, but further research is needed to establish best practices and standardized protocols for using LLMs in this role. In this paper, we contribute to this line of inquiry by exploring multiple LLM as a Judge configurations and comparing their performance against traditional metrics and human evaluations.

3. Methodology

3.1. Case Study

For this case study, we utilized internal documents provided by a consulting company as the foundation for building and evaluating a RAG system. The corpus consists of over 1,550 documents covering a wide range of topics relevant to the company's areas of expertise, including operations management, corporate finance, data science, supply chain optimization, organizational design, and agile transformation. These documents were indexed and stored in a vector database to support the retrieval component of the RAG pipeline. The primary use case for the system is to answer technical questions posed by consultants on these specialized topics.

Table 1. Evaluation Rubric: Response Accuracy and Helpfulness

Score	Description
1	The response is incomplete, factually incorrect, or irrelevant to the user's query, potentially leading to misunderstanding or misinformation.
2	The model attempts to answer but provides partially incorrect or vague information, with significant omissions or lack of clarity.
3	The model offers a generally accurate and relevant response, but may lack full detail or leave some aspects of the user's query unaddressed.
4	The response is factually sound and covers most key points with clarity, though there may be minor gaps in completeness or nuance.
5	The model delivers a comprehensive, precise, and clearly articulated answer that fully addresses the user's query with high factual integrity and helpful context.

To assess the performance of the two LLM-based QA systems integrated with RAG, we assembled a validation set comprising 50 real-world questions, all formulated in Portuguese. These questions mirror typical inquiries posed by consultants during their daily work. Each question was paired with a human-authored reference answer, created

by domain experts within the company who possess deep subject matter expertise. These experts ensured that the reference answers were both factually accurate and contextually aligned with the intended use case.

The same domain experts also evaluated the quality of the generated responses using two complementary methodologies: Direct Assessment and Pairwise Comparison. In the Direct Assessment setup, each answer was independently rated on a scale based on the criteria outlined in Table 1. In the Pairwise Comparison setup, experts were shown pairs of answers to the same question and asked to identify the superior one using the same evaluation criteria. Each evaluation task was assigned to a single expert.

By involving the original content creators and domain specialists in the evaluation process, we ensured that the annotations and judgments reflected real-world expectations and standards of quality. The combination of domain-specific data, expert annotation, and multiple evaluation protocols provides a robust foundation for assessing the alignment between human evaluation and automated metrics, including LLM as a Judge configurations. Figure 1 presents an example of a question.

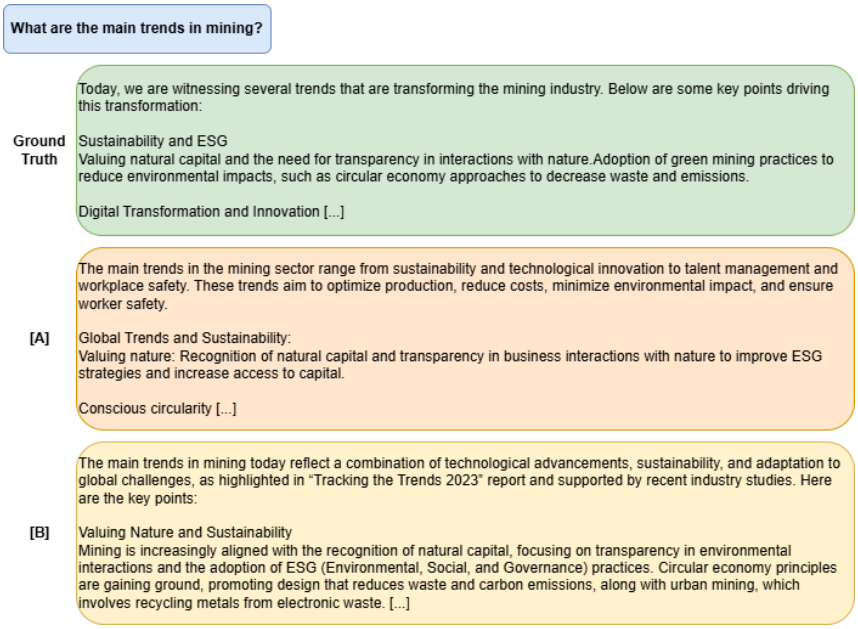

Figure 1. Illustrative example of an evaluation instance. The user query is presented alongside the human-annotated Ground Truth answer. Responses A and B represent the outputs from the two RAG systems compared in this study. This structure forms the basis for both human and LLM-based evaluation

3.2. Experiments

The experiments were conducted by comparing two distinct QA systems based on RAG. Both systems utilized the same vector database built with Chroma [Chroma 2025], employing a retriever configured with cosine similarity and a top-k value of 3. To construct the vector database, the original documents were first summarized using an LLM to reduce redundancy and enhance relevance, followed by the extraction of structured metadata to support more effective retrieval and filtering during inference.

The key difference between the two QA systems lies in the LLM used for answer generation. One system employed OpenAI's GPT-4o-mini [OpenAI 2025], a lightweight yet powerful variant of GPT-4 optimized for latency and cost efficiency. The other system used Google's Gemini 2.0 Flash [Google 2025], designed for high-throughput applications and also offering strong performance across various reasoning and language understanding benchmarks. While both models are optimized for speed and scalability, Gemini 2.0 Flash is generally known for its fast inference capabilities [Google 2025], whereas GPT-4o-mini offers broader compatibility with OpenAI's alignment and instruction-following standards [OpenAI 2025]. This comparison allowed us to evaluate model-specific differences in a controlled environment.

As traditional NLP metrics, we computed BLEU, ROUGE, METEOR, and BERTScore by comparing the model-generated responses against the human-annotated answers from our validation dataset. These metrics provide different perspectives on text quality: BLEU and ROUGE focus on lexical overlap [Papineni et al. 2002] [Lin 2004], METEOR incorporates stemming and synonym matching [Banerjee and Lavie 2005], and BERTScore leverages contextual embeddings to assess semantic similarity [Zhang et al. 2020].

To complement the traditional metrics, we adopted the LLM as a Judge paradigm using Prometheus 2 [Kim et al. 2024], a state-of-the-art fine-tuned LLM evaluation model. Prometheus 2 has demonstrated strong alignment with human preferences in various QA and summarization benchmarks, and is specifically trained to emulate human judgment across multiple evaluation dimensions [Kim et al. 2024]. In our experiments, Prometheus 2 was used in two settings: Pairwise Comparison and Direct Assessment. For both, we applied the criterion presented in Table 1.

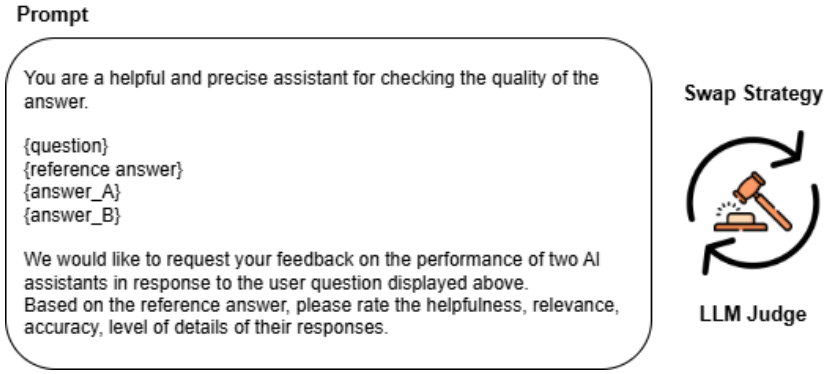

Figure 2. Illustration of the LLM-based judging setup using In-Context Learning. A prompt template is constructed with the original user question, a reference answer, and two candidate responses (Answer A and Answer B). The LLM is asked to evaluate both answers. To mitigate positional bias, a swap strategy is applied, where the positions of Answer A and B are alternated in separate evaluations, and results are aggregated. This setup allows for a scalable and consistent evaluation protocol using LLMs as automated judges.

In the Pairwise Comparison setup, Prometheus 2 was prompted with two model-generated answers and asked to choose the more helpful one. For Direct Assessment, it was instructed to rate a single answer on a 1 to 5 scale using the same evaluation criterion,

as presented in Table 1. This dual-mode application allowed us to analyze how different evaluation formats affect model judgment and their alignment with human ratings.

In addition to Prometheus 2, we explored an In-Context Learning approach inspired by JudgeLM [Zhu et al. 2025], a recent framework specifically designed to mitigate position bias and format bias in LLM-based evaluation. This method introduces a swap strategy, in which the order of candidate responses is randomly flipped during pairwise evaluation to reduce positional preference [Zhu et al. 2025]. By applying this technique, we aimed to improve the fairness and consistency of the LLM judge's outputs. Figure 2 presents the proposed approach.

All evaluation metrics, both traditional (BLEU, ROUGE, METEOR, BERTScore) and LLM-based (Prometheus 2, In-Context Learning), were benchmarked against the human annotations collected during the validation phase. This alignment analysis allowed us to assess the reliability of automated evaluation methods and their ability to replicate expert judgment in a domain-specific QA context. Figure 3 presents the evaluation methodology.

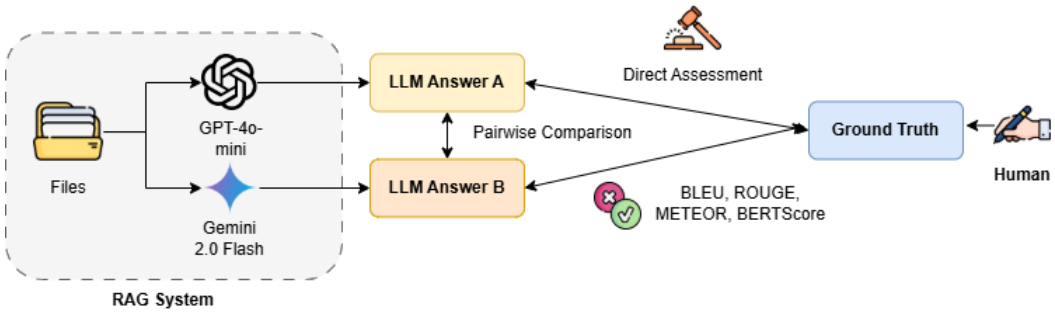

Figure 3. Overview of the evaluation methodology used for comparing two LLM backends (GPT-4o-mini and Gemini 2.0 Flash) within a Retrieval-Augmented Generation (RAG) system. Given the same input documents, each model produces a response. These outputs are then evaluated through both NLP metrics (BLEU, ROUGE, METEOR, and BERTScore) and human-aligned assessments. The human evaluations include Direct Assessment and Pairwise Comparison. Annotations are performed either by human raters or LLM-based judges.

Through this multifaceted evaluation strategy, our goal was to not only compare two modern RAG-based QA systems but also to investigate the extent to which different evaluation paradigms—ranging from classic metrics to advanced LLM judges—correlate with domain expert assessments. The combination of fine-tuned and in-context LLM evaluation provides valuable insights into the design of scalable, high-fidelity evaluation pipelines for real world applications.

4. Results

The initial phase of our results analysis focuses on the performance of the two RAG system configurations—GPT-4o-mini and Gemini 2.0 Flash—as measured by established traditional NLP evaluation metrics. Table 2 presents a comparative view of BLEU, ROUGE (specifically ROUGE-1, ROUGE-2, and ROUGE-L), METEOR, and BERTScore, all computed against our human-annotated validation dataset.

A prominent observation from Table 2 is that the absolute values obtained for BLEU, the various ROUGE scores, and METEOR are relatively modest for both models. This outcome is somewhat anticipated given the nature of these metrics, which heavily rely on n-gram overlap and lexical matches. In the context of RAG systems generating answers from a specialized corpus, responses are often abstractive, synthesizing information rather than merely extracting verbatim text. Consequently, even semantically accurate and helpful answers might exhibit low surface-level similarity with the ground truth answers, leading to low scores. This suggests that, from a purely lexical standpoint, the LLM-generated responses do not perfectly mirror the exact phrasing of the reference texts.

Table 2. Evaluation of RAG Systems using BLEU, ROUGE, METEOR, and BERTScore

Metric	GPT-4o-mini	Gemini 2.0 Flash
BLEU	0.073 ± 0.005	0.108 ± 0.006
ROUGE-1	0.406 ± 0.010	0.436 ± 0.009
ROUGE-2	0.177 ± 0.008	0.209 ± 0.007
ROUGE-L	0.207 ± 0.009	0.243 ± 0.010
METEOR	0.309 ± 0.006	0.332 ± 0.005
BERTScore	0.692 ± 0.004	0.709 ± 0.003

In contrast, BERTScore yielded considerably higher values across both models. This discrepancy is directly attributable to the fundamental differences in how these metrics operate. While BLEU, ROUGE, and METEOR are anchored in lexical similarity, BERTScore leverages contextual embeddings derived from pre-trained transformer models. This allows it to capture deeper semantic similarity, recognizing paraphrases and conceptually related terms even when there is little to no direct n-gram overlap between the predicted and reference answers. The notably higher BERTScore indicate a greater degree of semantic alignment between the LLM outputs and the ground truth, which is a more desirable characteristic for nuanced QA systems.

Despite the acknowledged limitations of traditional string-based metrics, particularly in capturing comprehensive answer quality, a consistent pattern emerges from the data in Table 2: the Gemini 2.0 Flash model consistently outperforms the GPT-4o-mini model across all evaluated metrics. This trend, observable in both lexical and semantic similarity scores, suggests that Gemini 2.0 Flash produces responses that are, on average, closer to the ground truth, whether assessed by word overlap or by underlying meaning. These initial quantitative results therefore lend support to the selection of Gemini 2.0 Flash as a more effective backend for our RAG system within this domain-specific context.

Subsequently, our investigation shifted towards evaluating the RAG systems through the lens of an LLM as a Judge, specifically employing the fine-tuned Prometheus 2 model. The initial approach utilized Direct Assessment. The comparative average scores assigned by Prometheus 2 and our human annotators for both GPT-4o-mini and Gemini 2.0 Flash are presented in Table 3. To quantify the alignment between these automated and human judgments, we calculated Pearson, Spearman, and Cohen's Kappa correlation coefficients. These yielded values of 0.254, 0.308, and 0.250, respectively. Such

low correlation coefficients strongly suggest a limited concordance between the scores assigned by Prometheus 2 in a Direct Assessment format and those provided by human experts. This indicates that relying solely on LLM as a Judge for absolute scoring via Direct Assessment may not be a consistently reliable method for capturing nuanced human perceptions of answer quality. Further supporting this, the direct agreement rate, where both judge and human assigned the same score, in this setup was a modest 66.7%.

Table 3. Comparison of Average Scores Assigned to RAG Models by Prometheus 2 and Human Judges

Evaluation Method	GPT-4o-mini	Gemini 2.0 Flash
LLM as a Judge (Prometheus 2)	*4.1*	*4.0*
Human Judgment	*3.5*	*4.5*

Given the limitations observed with Direct Assessment, we proceeded to apply a Pairwise Comparison methodology, still utilizing the Prometheus 2 evaluator. When Prometheus 2 was used in this pairwise setting, it determined that the Gemini 2.0 Flash model produced the better answer in 60.0% of the comparisons against GPT-4o-mini. Concurrently, human annotators performing the same pairwise task reported a win rate of 75.0% for Gemini 2.0 Flash. Crucially, the agreement rate between the LLM-based pairwise judgments and the human-based pairwise judgments reached 80.0%. This marked improvement in alignment confirms that Pairwise Comparison, when implemented with a capable fine-tuned LLM judge like Prometheus 2, offers a more human-congruent assessment methodology than Direct Assessment for these RAG systems.

To explore less computationally intensive alternatives to fine-tuned models, we also tested few-shot evaluators that leverage In-Context Learning. This approach, inspired by frameworks such as JudgeLM, involves providing the LLM judge with a few examples of question-answer pairs and their evaluations directly within the prompt, guiding its judgment without requiring model retraining. In this ICL setup, the LLM judge, using a general-purpose LLM, assigned the Gemini 2.0 Flash system a win rate of 73.3% in Pairwise Comparisons. However, despite this seemingly strong performance for Gemini, the agreement rate between this ICL-based judge and our human annotators dropped to 60.0%. This decrease highlights a potential inconsistency in alignment when moving from a specialized fine-tuned judge to a general-purpose LLM guided by only a few examples, suggesting that while ICL can achieve reasonable performance, its judgments may diverge more frequently from human consensus.

Collectively, these findings regarding LLM as a Judge configurations suggest a clear hierarchy in terms of human alignment. While In-Context Learning presents a more accessible and less computationally demanding alternative, its capacity to consistently replicate human judgment in our tests still lags behind that of specialized, fine-tuned evaluators, such as Prometheus 2. Furthermore, within the fine-tuned judge paradigm, Pairwise Comparison demonstrably yields superior alignment with human preferences compared to Direct Assessment. Consequently, for evaluation tasks demanding high reliability, interpretability, and close fidelity to human assessment, fine-tuning-based approaches,

particularly in a pairwise format, remain the preferable methodology for evaluating RAG QA systems.

5. Conclusions

This study embarked on a comprehensive evaluation of RAG-based QA systems, comparing traditional NLP metrics, various LLM as a Judge configurations, and human assessments on a domain-specific dataset. Our findings consistently demonstrated that Gemini 2.0 Flash outperformed GPT-4o-mini across traditional metrics, such as BLEU, ROUGE, METEOR, and particularly BERTScore, which proved more adept at capturing semantic similarity than purely lexical measures. More critically, our investigation into LLM as a Judge paradigms revealed that a fine-tuned model, Prometheus 2, when employed for Pairwise Comparison, achieved the highest human alignment with an 80.0% agreement rate. This significantly surpassed its performance in Direct Assessment (66.7% agreement) and also outperformed an In-Context Learning approach (60.0% agreement), highlighting the current superiority of fine-tuned, pairwise LLM judges for emulating human evaluation preferences.

The implications of these findings are twofold. Firstly, they contribute to the broader understanding of automated evaluation reliability, underscoring that while convenient, not all LLM as a Judge configurations offer the same level of fidelity to human judgment. The preference for fine-tuned, pairwise comparison suggests that tasks requiring high reliability and interpretability should prioritize such methods, despite their potentially higher setup cost compared to In-Context Learning or simpler traditional metrics. Secondly, they provide actionable insights for practitioners selecting LLM backends and evaluation methodologies for RAG QA systems, particularly in specialized domains

Future work should extend this comparative analysis to a wider array of LLM judges, including both open-source and proprietary models, and across diverse datasets and task types to assess the generalizability of our findings. Investigating more sophisticated In-Context Learning strategies, such as dynamic example selection or Chain of Thought prompting for judges, could also help bridge the alignment gap with fine-tuned models while retaining ICL's flexibility. Furthermore, exploring the root causes of disagreement between LLM judges and human evaluators, especially in cases of Direct Assessment and ICL, could lead to better prompting techniques or targeted fine-tuning datasets to enhance LLM judge capabilities.

Finally, research into the explainability of LLM judge decisions is necessary. Understanding why an LLM judge prefers one response over another, beyond a simple score or choice, will be crucial for building trust and enabling iterative improvement of both the QA systems and the evaluation frameworks themselves. Developing robust metrics to quantify the confidence and calibration of LLM judges, alongside methods for mitigating known biases such as verbosity or positional preference in more nuanced ways, will also be essential steps towards creating reliable and scalable automated evaluation pipelines for the next generation of generative AI systems.

References

Bai, Y., Ying, J., Cao, Y., Lv, X., He, Y., Wang, X., Yu, J., Zeng, K., Xiao, Y., Lyu, H., Zhang, J., Li, J., and Hou, L. (2023). Benchmarking foundation models with language-

model-as-an-examiner. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Goldstein, J., Lavie, A., Lin, C.-Y., and Voss, C., editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Chroma (2025). Chroma. Available on: `https://www.trychroma.com/`. Acessed in 18 May 2025.

Google (2025). Gemini 2.0 flash. Available on: `https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash?hl=pt-br`. Acessed in 18 May 2025.

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., Wang, Y., Gao, W., Ni, L., and Guo, J. (2024). A survey on llm-as-a-judge. *ArXiv*.

Ho, X., Huang, J., Boudin, F., and Aizawa, A. (2025). LLM-as-a-Judge: Reassessing the Performance of LLMs in Extractive QA. *arXiv preprint arXiv:2504.11972*.

Kamalloo, E., Dziri, N., Clarke, C., and Rafiei, D. (2023). Evaluating open-domain question answering in the era of large language models. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada. Association for Computational Linguistics.

Kim, S., Suk, J., Longpre, S., Lin, B. Y., Shin, J., Welleck, S., Neubig, G., Lee, M., Lee, K., and Seo, M. (2024). Prometheus 2: An open source language model specialized in evaluating other language models. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.

Ko, M., Lee, J., Kim, H., Kim, G., and Kang, J. (2020). Look at the first sentence: Position bias in question answering. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1109–1121, Online. Association for Computational Linguistics.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

OpenAI (2025). Gpt-4o mini: advancing cost-efficient intelligence. Available on: `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`. Acessed in 18 May 2025.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on As-*

sociation for Computational Linguistics, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Qin, Z., Jagerman, R., Hui, K., Zhuang, H., Wu, J., Yan, L., Shen, J., Liu, T., Liu, J., Metzler, D., Wang, X., and Bendersky, M. (2024). Large language models are effective text rankers with pairwise ranking prompting. In Duh, K., Gomez, H., and Bethard, S., editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1504–1518, Mexico City, Mexico. Association for Computational Linguistics.

Sai, A. B., Mohankumar, A. K., and Khapra, M. M. (2022). A survey of evaluation metrics used for nlg systems. *ACM Comput. Surv.*, 55(2).

Schluter, N. (2017). The limits of automatic summarisation according to ROUGE. In Lapata, M., Blunsom, P., and Koller, A., editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 41–45, Valencia, Spain. Association for Computational Linguistics.

Wang, Y., Yu, Z., Zeng, Z., Yang, L., Wang, C., Chen, H., Jiang, C., Xie, R., Wang, J., Xie, X., Ye, W., Zhang, S.-B., and Zhang, Y. (2023). Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *ArXiv*, abs/2306.05087.

Yu, Q., Zheng, Z., Song, S., Li, Z., Xiong, F., Tang, B., and Chen, D. (2024). xfinder: Large language models as automated evaluators for reliable evaluation. In *International Conference on Learning Representations*.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Zhu, L., Wang, X., and Wang, X. (2025). Judgelm: Fine-tuned large language models are scalable judges. *International Conference on Learning Representations*.