Validating ESG Commitments with RAG-Enhanced Large Language Models: Toward Transparent and Reliable Sustainability Disclosure

Hsin-Ting Lu, Min-Yuh Day*

Graduate Institute of Information Management
National Taipei University, New Taipei, Taiwan
E-mail: hsintinglubob@gmail.com; myday@gm.ntpu.edu.tw

Abstract

Corporate sustainability reports are central to sustainable governance and are widely used as an indicator of corporate responsibility. Yet, ESG commitments often contain vague or unverifiable statements, undermining transparency and heightening the risk of greenwashing. Despite increasing global emphasis on ESG disclosure, current approaches still lack structured and comparable mechanisms for verifying corporate commitments.

This study integrates Retrieval-Augmented Generation (RAG) with large language models (LLMs), focusing on the French subset of the ML-Promise dataset to explore automated ESG commitment validation. The framework addresses four subtasks: Promise Status, Evidence Status, Evidence Quality (clarity of supporting evidence), and Verification Timeline. Results show that while LLMs struggle with ambiguous commitments and insufficient evidence, RAG enhances performance, particularly for reasoning-intensive tasks such as Evidence Quality and Verification Timeline.

The proposed framework provides both methodological insights for applying LLMs to ESG reporting and practical implications for regulators and corporations, offering a language-specific benchmark to enhance transparency and accountability in sustainability disclosure.

Keywords:
Retrieval-Augmented Generation (RAG), Large Language Models (LLMs), ESG Reports, ESG Commitment Validation, Greenwashing Detection

1. Introduction

Corporate sustainability reports have become an important channel for companies to communicate their Environmental, Social, and Governance (ESG) commitments. However, many of these commitments remain vague, difficult to verify, or selectively presented, raising concerns about greenwashing, which refers to the practice of exaggerating environmental efforts in reports without meeting regulatory standards [1]. To address this challenge, this study adopted the French subset of the ML-Promise dataset (~400 samples) [2] as the core corpus and focuses on four subtasks: Promise Identification, Supporting Evidence Assessment, Evidence Quality, and Timing for Verification. This study investigated how Retrieval-Augmented Generation (RAG) can be integrated with large language models (LLMs) of different scales, leveraging retrieval to supplement external knowledge and improve classification and reasoning accuracy.

Accordingly, this study addressed the following research questions: (1) Can RAG significantly improve LLM performance in ESG promise verification tasks compared with non-RAG baselines? (2) Do RAG-enhanced LLMs show different performance across the four subtasks (Promise Identification, Supporting Evidence Assessment, Evidence Quality, and Timing for Verification)? (3) How does model scale (large, medium, small) affect the effectiveness of RAG in ESG verification, and can smaller models benefit from retrieval to close the gap with larger models?

In terms of methodology, this study adopted a RAG-enhanced framework in which ESG promises and retrieved contexts are fed into large, medium, and small LLMs. The evaluation relies on F1-score as the main metric to address class imbalance issues. Preliminary results suggest that RAG can significantly improve performance in higher-level reasoning tasks, such as evidence quality and timeline prediction, while smaller models benefit substantially from retrieval, reducing the performance gap with larger models.

The contributions of this study lie in providing empirical analysis for monolingual ESG verification (French subset), while illustrating the performance differences of retrieval augmentation across models of different scales. From a managerial perspective, the proposed approach can help regulators efficiently identify unsupported or exaggerated corporate commitments and guide companies to enhance the verifiability and credibility of their disclosures, thereby strengthening investor and public trust in sustainability reports.

2. Literature Review

2.1 ESG Reporting and the Challenge of Greenwashing

Sustainability reports have become an important reference for assessing corporate performance in environmental, social, and governance (ESG) dimensions, as well as a key channel for firms to communicate commitments and demonstrate accountability. As ESG disclosure gains increasing attention, some companies selectively reveal information to create a positive image and attract stakeholder support, rather than disclosing potential negative environmental impacts. This practice has led to the emergence of greenwashing incidents. Recent studies have attempted to detect such behavior. For example, [3] introduced the A3CG dataset as a novel benchmark to enhance the robustness of ESG analysis under

greenwashing contexts. Similarly, [4] fine-tuned the ClimateBERT model to improve accuracy in identifying misleading disclosures.

Although NLP-based ESG analysis methods provide valuable insights from sustainability reports, they still fall short in explaining the credibility of corporate claims. Furthermore, their effectiveness in handling multilingual verification tasks remains limited, indicating the need for more systematic and scalable solutions.

2.2 Large Language Models: Capabilities and Scalability

In recent years, LLMs have advanced rapidly, achieving remarkable performance across a wide range of natural language processing (NLP) tasks, particularly in text generation [5] . A key feature of LLMs is their scalability: large models, often with hundreds of billions of parameters, demonstrate superior capabilities in complex reasoning and cross-domain transfer tasks but come with extremely high computational and financial costs. In contrast, small- and medium-scale models are more efficient and easier to deploy, though their performance is often limited without external augmentation. Recent studies show that Retrieval-Augmented Generation (RAG) can effectively help smaller models narrow the performance gap with larger ones, making them more practical for real-world applications [6] . Nevertheless, LLMs continue to face challenges, with one of the most prominent being hallucination, which refers to the generation of fluent yet factually incorrect content that undermines reliability [7] . Furthermore, the training and deployment costs of large-scale models remain prohibitively high, limiting their accessibility in multilingual and domain-specific applications. These constraints underscore the necessity of systematically comparing models of different scales, especially when combined with RAG, to better balance performance and efficiency.

2.3 Retrieval-Augmented Generation for Knowledge-Intensive Tasks

Although LLMs achieve strong performance across many natural language processing (NLP) tasks, they remain constrained by hallucinations and by limited access to up-to-date or domain-specific knowledge[8]. These limitations are especially salient in knowledge-intensive settings such as fact verification and open-domain question answering. Retrieval-Augmented Generation (RAG) has been proposed to mitigate this problem: a retriever is coupled with a generator so that the model first retrieves relevant passages from external corpora and then uses them as supplemental context to support generation [9] . This hybrid architecture improves factuality and interpretability because generated content can be traced back to retrieved sources. Recent studies further show that RAG strategies can significantly enhance model performance, leading to steady gains on complex reasoning and knowledge-intensive tasks [10] . RAG has been widely applied to open-domain QA, multi-hop reasoning, and specialized text analytics such as clinical trial data analysis [11] and legal document processing. Its core value lies in improving factual reliability and task verifiability. Nevertheless, RAG's effectiveness remains highly dependent on retrieval quality and corpus coverage, underscoring the need for systematic evaluation and validation in emerging applications such as ESG promise verification.

2.4 The ML-Promise Dataset for Multilingual ESG Commitment Verification

ML-Promise is the first multilingual dataset specifically designed for corporate promise verification, covering English, French, Chinese, Japanese, and Korean, with approximately 3,010 samples collected from ESG reports across five countries. The dataset was developed to address the challenges of evaluating corporate sustainability commitments, particularly in response to cases where companies employ misleading information to create an overly positive environmental image, a practice commonly referred to as greenwashing. ML-Promise organizes the verification task into four subtasks: (1) Promise Identification, (2) Supporting Evidence Assessment, (3) Evidence Quality, and (4) Timing for Verification. In this study, we focus on the French subset (~400 samples) to investigate how Retrieval-Augmented Generation (RAG) combined with large, medium, and small LLMs can improve performance in promise verification tasks [2] .

3. Research Methodology

3.1 System Architecture

Figure 1 illustrates the overall experimental framework of this study. This study used the ML-Promise French subset as the primary dataset. Under the RAG setting, the multilingual-e5-base retriever selects the top-6 most relevant training examples, which are incorporated as auxiliary context for the model. This study compared three language models at different scales, namely Gemma3-4B, Gemma3-12B, and GPT-5, which represent small, medium, and large configurations, respectively. All models are evaluated on four subtasks: Promise Status, Evidence Status, Evidence Quality, and Verification Timeline. Performance is measured using Macro-F1 as the primary metric to account for class imbalance and enable a comprehensive comparison.

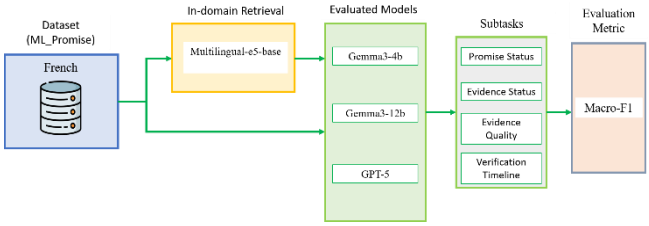

Figure 1. Proposed research workflow for ESG promise verification

3.2 Dataset

This study adopted ML-Promise [2] and focuses on the French subset. The dataset is drawn from corporate ESG disclosures. Our evaluation is conducted exclusively on the French test set (n = 400). The training split is not used for supervised fine-tuning, but instead serves solely as the

retrieval corpus for the RAG component, providing candidate passages during inference. Each sample is annotated for four single-label subtasks:

- Promise Status: whether a concrete or organization-level commitment is present (Yes / No).
- Evidence Status: whether verifiable supporting evidence is provided (Yes / No).
- Evidence Quality: clarity of the evidence (Clear, Not Clear, Misleading, N/A).
- Verification Timeline: expected timeframe for fulfilling the commitment (Already, Less than 2 years, 2 to 5 years, More than 5 years, N/A).

To illustrate dataset composition and potential imbalance, Figure 2 presents the label distribution in the French test set (n = 400), while the training split used for retrieval is not shown, since it is not part of evaluation.

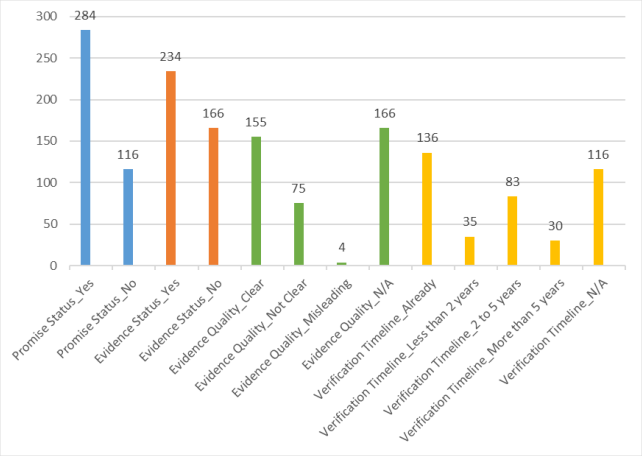

Figure 2. Label distribution of the French test set (n = 400), which is used for evaluation in this study

3.3 Model Selection

In this study, we evaluated three language models spanning small–large scales: Gemma 3: 4B (small), Gemma 3: 12B (medium), and GPT-5 (large). The 4B model offers a lightweight configuration suitable for resource-constrained and latency-sensitive scenarios; the 12B model provides stronger reasoning and contextual understanding as an effective mid-scale option; and GPT-5 serves as a state-of-the-art large model. This stratified selection enables us to systematically examine how Retrieval-Augmented Generation (RAG) interacts with model scale on the ML-Promise French subset, and whether retrieval allows small and medium models to narrow the gap to a large model under comparable Macro-F1 evaluation.

3.4 Retrieval Corpus and Indexing

For the RAG component, we built a retrieval corpus using the French training split of ML-Promise. Importantly, this split was not used for supervised fine-tuning but served solely as the retrieval source. Each training sample was segmented and encoded using the multilingual-e5-base model to construct a FAISS index.

During inference, the system retrieved the top-6 passages from this index for each test instance, which were then appended to the prompt as contextual evidence. To ensure data integrity and avoid test leakage, we applied the following measures:

- Strict separation of splits – only training samples were indexed; the French test set (n = 400) was never included in the retrieval corpus.
- Near-duplicate removal – we performed string hashing and similarity checks to ensure that no near-identical text fragments from the test set existed in the retrieval index.
- Retrieval purpose – the indexed passages were used exclusively to provide auxiliary context; model outputs were always evaluated only against the test set labels.

This design ensures that RAG performance reflects contextual augmentation rather than inadvertent exposure to test content.

3.5 Evaluation Metric

This study adopted the macro-averaged F1 score (Macro-F1) as the sole evaluation metric. Macro-F1 calculates the F1 score for each class independently and then takes the unweighted average, ensuring equal importance for both majority and minority classes. This property is crucial for the ML-Promise dataset, where the distribution across subtasks (Promise Identification, Supporting Evidence Assessment, Evidence Quality, and Timing for Verification) is imbalanced. Compared with accuracy, which may be biased toward majority classes, Macro-F1 provides a fairer and more reliable assessment of classification and reasoning performance, especially when evaluating how Retrieval-Augmented Generation (RAG) enhances LLMs of different scales.

4. Experiment Results and Analysis

4.1 Overall Results with Baseline

Table 1 presents the overall experimental results for French ESG promise verification across four subtasks. RAG improves performance in most cases, with notable gains in Evidence Quality and Verification Timeline, which require higher-level reasoning. However, Supporting Evidence shows slight declines for Gemma3-4B and GPT-5, indicating task-dependent sensitivity to retrieval quality. Among the models, GPT-5 achieves competitive results, while Gemma3-12B also performs strongly. The ML-Promise baseline was reported with GPT-4o under the dataset's multilingual setting. Although GPT-5 demonstrates stronger reasoning capabilities, it does not consistently surpass GPT-4o, likely due to French data optimization, the small subset size (~400 samples), and evidence-task sensitivity. Nevertheless, GPT-5 remains competitive overall, especially with RAG, narrowing the gap with the reported baseline.

Table 1. Overall Experimental Results on French ESG Promise Verification (Macro-F1), with Comparisons to ML-Promise Baseline

RAG Setting	Task	Gemma3-4B	Gemma3-12B	GPT-5	ML_Promise French Dataset
w/o RAG	Promise Identification	0.509	0.734	0.687	**0.816**
	Supporting Evidence Assessment	**0.573**	0.528	**0.787**	**0.746**
	Evidence Quality	0.238	0.269	0.365	0.443
	Timing for Verification	0.211	**0.422**	0.418	0.523
w/ RAG	Promise Identification	**0.625**	**0.754**	**0.756**	0.798
	Supporting Evidence Assessment	0.523	**0.666**	0.749	0.732
	Evidence Quality	**0.285**	**0.330**	**0.419**	**0.487**
	Timing for Verification	**0.301**	0.411	**0.420**	**0.601**

4.2 Subtask-Level Performance Analysis

Figures 3–6 present subtask-level comparisons of w/ vs. w/o RAG across models. For Promise Identification (Figure 3), RAG yields moderate gains, especially for Gemma3-4B (+0.116, +22.8%), while improvements for larger models are smaller. Supporting Evidence Assessment (Figure 4) shows mixed results: Gemma3-12B improves substantially (+0.138, +26.1%), but GPT-5 (–0.038) and Gemma3-4B (–0.050) slightly decline, suggesting retrieval quality critically affects this task. In evidence quality (Figure 5), all models benefit, with gains ranging from +0.047 to +0.061, confirming RAG's strength in enhancing nuanced reasoning over promise–evidence pairs. For Verification Timeline (Figure 6), the largest benefit is observed in Gemma3-4B (+0.090, +42.7%), while larger models show minimal changes. Overall, RAG proves most effective in reasoning-intensive subtasks (evidence quality and Timeline), whereas Supporting Evidence remains highly sensitive to retrieval noise.

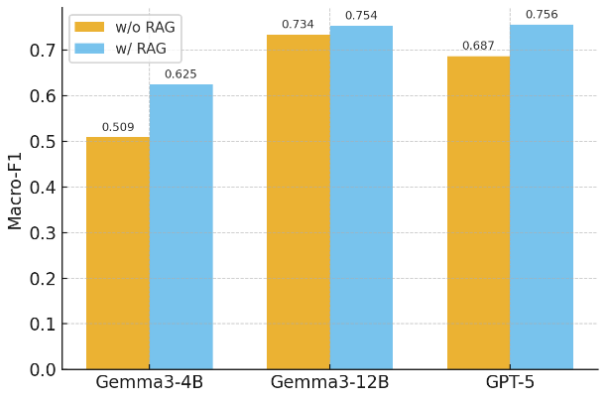

Figure 3. Subtask-level performance on Promise Identification (w/ vs. w/o RAG across models).

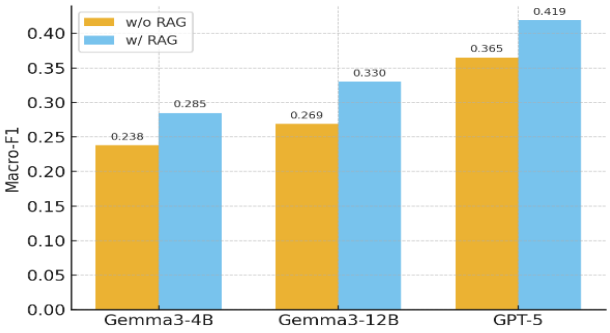

Figure 4. Subtask-level performance on Supporting Evidence Assessment (w/ vs. w/o RAG across models).

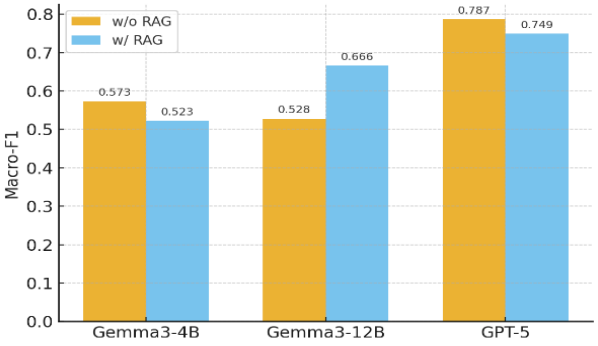

Figure 5. Subtask-level performance on evidence quality of the Promise–Evidence Pair (w/ vs. w/o RAG across models).

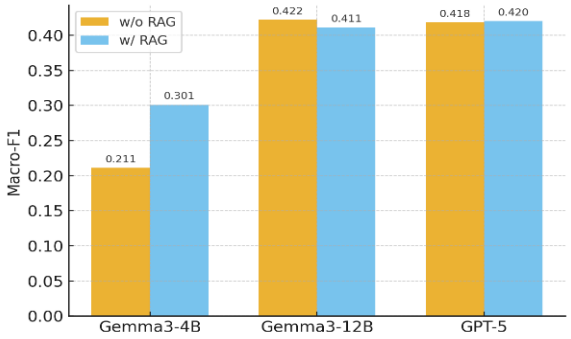

Figure 6. Subtask-level performance on Timing for Verification (w/ vs. w/o RAG across models).

4.3 Effect of Model Scale

Table 2 summarizes the subtask-level results with RAG across small (Gemma3-4B), medium (Gemma3-12B), and large (GPT-5) models, with ΔF1 indicating relative improvements over the no-RAG setting. The results reveal several patterns. For Promise Identification, GPT-5+RAG achieves the best performance (0.756), although the largest relative gain is observed in Gemma3-4B (+0.116). In Supporting Evidence, Gemma3-12B+RAG performs best (0.666) with a substantial gain (+0.138), while GPT-5 slightly declines (–0.038). For evidence quality, GPT-5+RAG achieves the highest score (0.419) with consistent gains across all models (+0.047 to +0.061). Finally, in Verification Timeline, GPT-5+RAG again leads in absolute performance (0.420), but Gemma3-4B shows the largest relative improvement (+0.090).

Overall, these findings suggest that RAG helps small models achieve notable relative improvements, though their absolute scores remain lower than medium and large models. Medium-scale Gemma3-12B benefits most in Supporting Evidence, effectively narrowing the gap with GPT-5. Large-scale GPT-5 exhibits the strongest absolute performance, but its relative improvements are modest, reflecting its strong baseline capabilities. Taken together, RAG proves most valuable for small and medium models, enabling them to close part of the performance gap with large-scale LLMs.

Table 2. Subtask-level Macro-F1 with RAG across small (Gemma3-4B), medium (Gemma3-12B), and large (GPT-5) models, with ΔF1 relative to no-RAG baseline. Bold values indicate the best performance per subtask.

Task	Gemma3-4B (ΔF1)	Gemma3-12B (ΔF1)	GPT-5(ΔF1)
Promise Identification	0.625 (+0.116)	0.754 (+0.020)	**0.756** (+0.069)
Supporting Evidence	0.523 (–0.050)	0.666 (+0.138)	**0.749** (–0.038)
Evidence Quality	0.285 (+0.047)	0.330 (+0.061)	**0.419** (+0.054)
Verification Timeline	0.301 (+0.090)	0.411 (–0.011)	**0.420** (+0.002)

5. Conclusion

This study examined the impact of Retrieval-Augmented Generation (RAG) on ESG commitment verification using the French subset of the ML-Promise dataset. Findings show that RAG consistently improves overall performance compared with non-RAG baselines, with the most notable gains in evidence quality and verification timeline. Effects, however, vary across subtasks: RAG yields strong improvements in promise identification and supporting evidence, though the latter remains sensitive to retrieval quality. Model scale further moderates these benefits, as small and medium models achieve the largest relative improvements, narrowing the gap with large models, while GPT-5 maintains the strongest absolute performance but gains only marginally from retrieval.

In terms of research contributions, this study introduced a RAG-enhanced framework that balances efficiency and accuracy in monolingual ESG verification (French subset), filling a gap in prior work that has lacked systematic, evidence-based evaluation. Future research will extend this framework to multilingual corpora. This framework advances methodological rigor by providing a scalable approach for assessing corporate commitments across diverse languages and contexts.

In terms of managerial implications, the proposed approach can assist regulators and stakeholders in identifying vague or unsupported commitments, thereby improving the credibility and transparency of sustainability disclosures and encouraging firms to fulfill their ESG promises more rigorously. Future research will extend the framework to larger multilingual corpora, optimize retrieval strategies to reduce noise, and explore applicability to other languages and domain-specific ESG contexts.

Acknowledgments

This research was supported by the Industrial Technology Research Institute (ITRI) and National Taipei University (NTPU), Taiwan, under grants NTPU-114A513E01 and NTPU-113A513E01; the National Science and Technology Council (NSTC), Taiwan, under grant NSTC 114-2425-H-305-003-; and National Taipei University (NTPU) under grant 114-NTPU_ORDA-F-004.

References

[1] C. Xu, Y. Miao, Y. Xiao, and C. Lin, "DeepGreen: Effective LLM-Driven Green-washing Monitoring System Designed for Empirical Testing--Evidence from China," *arXiv preprint arXiv:2504.07733,* 2025.

[2] Y. Seki, H. Shu, A. Lhuissier, H. Lee, J. Kang, M.-Y. Day, and C.-C. Chen, "ML-Promise: A Multilingual Dataset for Corporate Promise Verification," *arXiv preprint arXiv:2411.04473,* 2024.

[3] K. Ong, R. Mao, D. Varshney, E. Cambria, and G. Mengaldo, "Towards Robust ESG Analysis

Against Greenwashing Risks: Aspect-Action Analysis with Cross-Category Generalization," *arXiv preprint arXiv:2502.15821,* 2025.

[4] A. Vinella, M. Capetz, R. Pattichis, C. Chance, and R. Ghosh, "Leveraging language models to detect greenwashing," *arXiv preprint arXiv:2311.01469,* 2023.

[5] Y. Xie, C. Wang, J. Yan, J. Zhou, F. Deng, and J. Huang, "Making small language models better multi-task learners with mixture-of-task-adapters," in *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 2024, pp. 1094-1097.

[6] K. Li, L. Zhang, Y. Jiang, P. Xie, F. Huang, S. Wang, and M. Cheng, "LaRA: Benchmarking Retrieval-Augmented Generation and Long-Context LLMs--No Silver Bullet for LC or RAG Routing," *arXiv preprint arXiv:2502.09977,* 2025.

[7] X. Lin, Y. Ning, J. Zhang, Y. Dong, Y. Liu, Y. Wu, X. Qi, N. Sun, Y. Shang, and P. Cao, "LLM-based Agents Suffer from Hallucinations: A Survey of Taxonomy, Methods, and Directions," *arXiv preprint arXiv:2509.18970,* 2025.

[8] J. Wallat, M. Heuss, M. D. Rijke, and A. Anand, "Correctness is not Faithfulness in Retrieval Augmented Generation Attributions," in *ICTIR 2025 - Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval*, 2025, pp. 22-32, doi: 10.1145/3731120.3744592. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-105013792040&doi=10.1145%2f3731120.3744592&partnerID=40&md5=dbd1f068b642483b10c28d3e9921b088

[9] L. Zhang, Z. Jiang, H. Chi, H. Chen, M. Elkoumy, F. Wang, Q. Wu, Z. Zhou, S. Pan, and S. Wang, "Diagnosing and Addressing Pitfalls in KG-RAG Datasets: Toward More Reliable Benchmarking," *arXiv preprint arXiv:2505.23495,* 2025.

[10] S. Krishna, K. Krishna, A. Mohananey, S. Schwarcz, A. Stambler, S. Upadhyay, and M. Faruqui, "Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation," *arXiv preprint arXiv:2409.12941,* 2024.

[11] Y. Zheng, B. Li, Z. Lin, Y. Luo, X. Zhou, C. Lin, G. Li, and J. Su, "Revolutionizing Database Q&A with Large Language Models: Comprehensive Benchmark and Evaluation," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2025, vol. 2, pp. 5960-5971, doi: 10.1145/3711896.3737405. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-105014326087&doi=10.1145%2f3711896.3737405&partnerID=40&md5=57754945496427b3f9fda76ba90da80c