Large Language Models in Materials Science: Assessing RAG Evaluation Frameworks through graphene synthesis

Zen Han Cho[*1] Leonard Wei Tat Ng[*1]

[*]*Equal contribution* [1]*[School of Materials Science and Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798]*

Correspondence to: Leonard Wei Tat Ng leonard.ngwt@ntu.edu.sg

1. Introduction

Large language models (LLMs) and retrieval-augmented generation (RAG)[1] systems are increasingly applied to scientific research, yet their evaluation in specialized domains remains poorly validated. Existing benchmarks such as GPQA [2] and MaScQA [3] primarily rely on structured question formats (e.g., multiple-choice or predefined numerical answers), which limits their ability to evaluate open-ended scientific reasoning. Here, using graphene synthesis as a case study, we present the first systematic assessment of open-ended automated RAG evaluation frameworks against human expert judgment, revealing both their sensitivity to retrieval benefits and their fundamental limitations in scientific contexts.

2. Methodology

A domain-specific RAG pipeline was constructed using 300 peer-reviewed articles on graphene synthesis, from which synthesis methodologies were extracted, standardized, and embedded into a vector database for semantic retrieval. Twenty domain-specific questions spanning across major fabrication methods [4], synthesis of graphene derivatives, application-driven synthesis strategies, and mechanistic understanding were developed by a materials science expert, with ground truth answers established through a rigorous double-blind consensus process involving three independent domain experts. Experts generated answers based solely on their professional knowledge, without access to the retrieval corpus, ensuring that ground truth reflected genuine scientific understanding rather than corpus-specific artifacts.

During response generation, two inference modes were employed: RAG and standard non-RAG inference. In the RAG setting, each query was embedded and matched against the vector database using cosine similarity, with the top five most relevant papers retrieved at inference time and combined into a structured prompt to ground answer generation in domain-specific literature. In the non-RAG setting, responses were generated solely from the query without access to external context. Both modes were applied to two LLMs: Gemini-2.5-Flash, a proprietary closed-source model, and Qwen2.5-7B, an open-source model, enabling analysis of retrieval benefits across model classes.

Model outputs were evaluated using four assessment approaches: RAGAS [5], BERTScore [6], an LLM-as-a-judge, and blinded expert human evaluation. To enable consistent comparison across all evaluators, factual correctness (FC) was used as the primary evaluation metric. Within RAGAS, the FC metric assesses overlap between generated responses and ground truth using claim-level decomposition. BERTScore was employed as an alternative method for evaluating semantic alignment between generated answers and ground truths. The LLM-as-a-judge approach used GPT-4o with a structured evaluation rubric to assign FC scores based on factual alignment with the ground truth. In parallel, nine expert human evaluators independently scored responses using the same rubric under blinded conditions. This evaluation design ensured direct comparability of factual correctness assessments across automated and human evaluators.

3. Results

Human evaluation revealed clear performance hierarchies demonstrating the value of retrieval augmentation for scientific applications. RAG-Gemini achieved the highest average score (6.92), followed by RAG-Qwen (6.68), standard Gemini (6.37), and standard Qwen (5.68). The performance improvements from retrieval augmentation were substantial: 0.55 points for Gemini and 1.00 points for Qwen, representing 9% and 17% relative improvements respectively.

Notably, the impact of retrieval was more pronounced for smaller open-source models: RAG-Qwen not only exceeded the performance of standard Gemini despite its smaller size (7B vs. Gemini's larger scale) but also showed nearly twice the relative gain in FC compared to Gemini-2.5-Flash. This demonstrates that retrieval enhances smaller open-source models to the point where they can compete effectively with

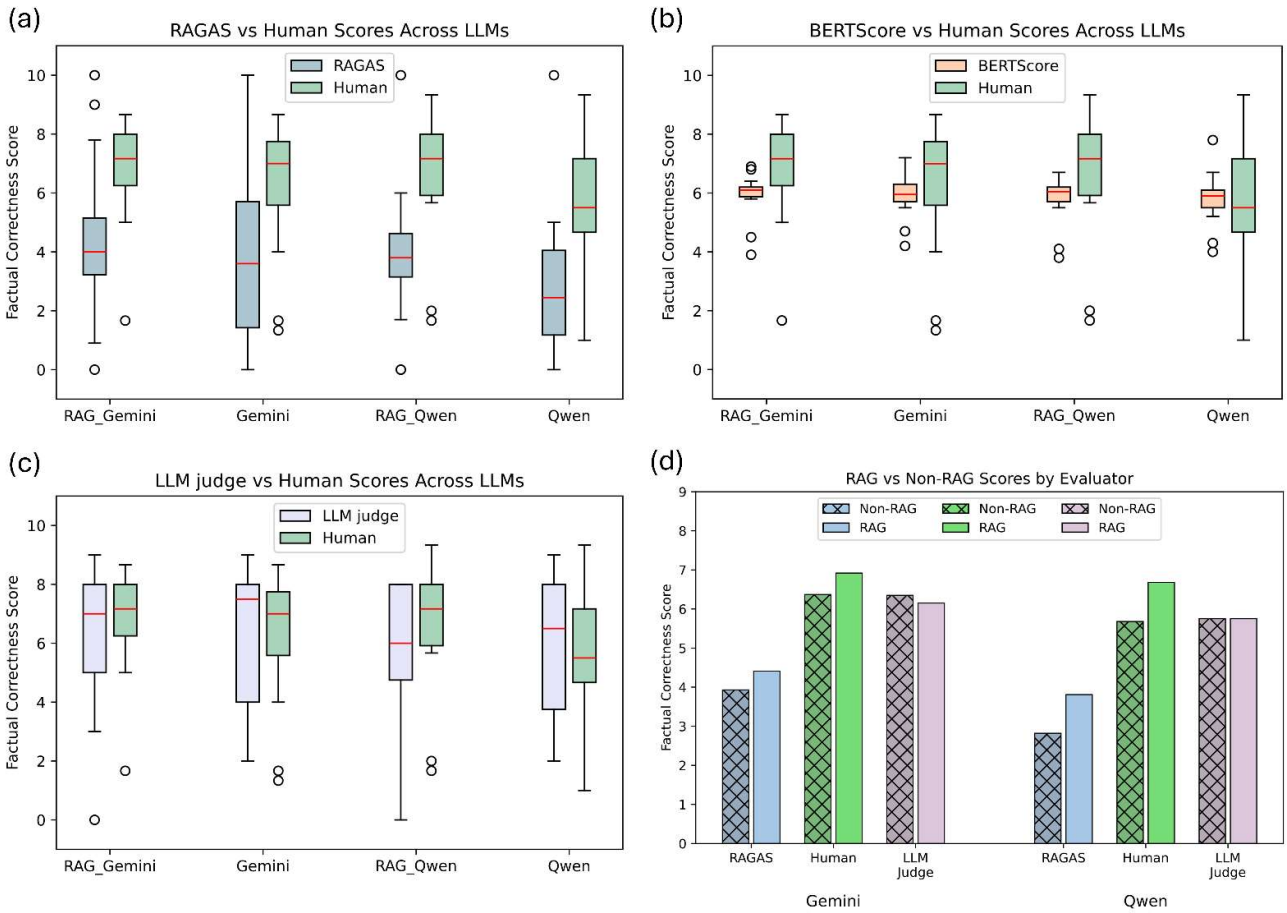

Figure 1: Comparison of factual correctness scores across LLMs and evaluators. (a) RAGAS vs. human scores for RAG and non-RAG variants of Gemini and Qwen. RAGAS underestimates factual correctness scores and shows poor alignment with human evaluation. (b) BERTScore vs. human scores across same LLMs. BERTScore show low variance in scores. (c) LLM Judge vs. human scores, showing the closest alignment in both distribution and median values. (d) Mean factual correctness scores across RAG and non-RAG LLMs by evaluator. Human and RAGAS reflect consistent factual gains from RAG augmentation, unlike LLM judge.

larger proprietary alternatives in domain-specific applications.

The relationship between automated evaluators and human judgment reveals critical insights about evaluation framework reliability in scientific contexts. RAGAS exhibited the largest absolute deviation from human scores (73.5% average difference) yet demonstrated the highest sensitivity to retrieval-augmented performance improvements (Figure 1a). RAGAS successfully captured the relative performance gains observed by human evaluators: 0.52-point improvement for Gemini (vs. 0.55 human-observed) and 1.03-point improvement for Qwen (vs. 1.00 human-observed).

BERTScore showed minimal absolute deviation (8.78%) but suffered from restricted score distribution ($\sigma = 0.70$), clustering most outputs between 5.19-6.59 (Figure 1b). When applied to human evaluation patterns, only 24% of human scores fell within BERTScore's expected range, indicating poor alignment with human scoring patterns despite superficial agreement in average scores. Consequently, BERTScore lacks the interpretability and responsiveness needed for evaluating factual correctness.

LLM judge demonstrated both low absolute deviation (7.53%) and appropriate score distribution ($\sigma = 2.24$), providing the closest overall alignment with human evaluation patterns (Figure 1c). However, LLM judge failed to capture retrieval augmentation benefits consistently, incorrectly favouring standard Gemini over RAG-Gemini and showing minimal differentiation between Qwen variants (Figure 1d)

References

[1] P. Zhao *et al.*, "Retrieval-Augmented Generation for AI-Generated Content: A Survey," Feb. 2024, [Online]. Available: http://arxiv.org/abs/2402.19473

[2] D. Rein *et al.*, “GPQA: A Graduate-Level Google-Proof Q&A Benchmark,” Nov. 2023, [Online]. Available: http://arxiv.org/abs/2311.12022

[3] M. Zaki, Jayadeva, Mausam, and N. M. A. Krishnan, “MaScQA: investigating materials science knowledge of large language models,” *Digital Discovery*, vol. 3, no. 2, pp. 313–327, 2024, doi: 10.1039/D3DD00188A.

[4] R. S. Perala, N. Chandrasekar, R. Balaji, P. S. Alexander, N. Z. N. Humaidi, and M. T. Hwang, “A comprehensive review on graphene-based materials: From synthesis to contemporary sensor applications,” Jun. 01, 2024, *Elsevier Ltd*. doi: 10.1016/j.mser.2024.100805.

[5] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, “Ragas: Automated Evaluation of Retrieval Augmented Generation,” Apr. 2025, [Online]. Available: http://arxiv.org/abs/2309.15217

[6] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTScore: Evaluating Text Generation with BERT,” Apr. 2019, [Online]. Available: http://arxiv.org/abs/1904.09675