Bridging the Gap: Enhancing and Evaluating Zero-Cost LLMs for Scientific Question Answering

Mahule Roy[*]
Institute of Biomedical Engineering
University of Oxford, UK
mahule.roy@kellogg.ox.ac.uk

Snehal Rakshit
IIT Kharagpur

Shinjini Mondal
IISER Pune

Rahul Jaisy
AEC, Guwahati

Varun Ramanathan Alagappan
IISER Tirupati

Eishaan Khatri
RGIPT Jais

Vedant Vijay Patil
IIT Kharagpur

Archishman Banerjee
St. Xavier's College, Kolkata

Mariane Desiree C. Avendaño
TU Delft

Yashika Sharma
GGSIPU (MAIT),Delhi

Elumalai Oviya Dharshini
Independent Researcher

Nanditha Immidi
IACS Kolkata

Theo Fraser
The Open University

Emanuel Espinoza Prado
UCR

Gayathri Aishwarya E
ISRO

Akash Kanji
Jadavpur University

Abstract

Large language models that are accessible at zero monetary cost, either on open-source platforms or public APIs, have broad accessibility, but lack factual accuracy and domain-level reasoning in scientific applications. We compare zero-cost access LLMs with a specially designed benchmark of validated questions in materials science and physics, and we find significant weaknesses in zero-shot performance. To combat this, we present an improvement pipeline that integrates dense retrieval, structured grounding through a scientific knowledge graph, and unsupervised trend analysis from live literature (e.g., arXiv). The system feeds both an unstructured and structured context to the LLM with the aid of sophisticated prompting methods. We introduce three domain-relevant metrics —**Factual Correctness in Physical Sciences (FC-PS)**, **Quantified Physicochemical Constraints Score (QPCS)**, and **Literature Alignment** — to analyze outputs beyond typical NLP benchmarks. Our approach reduces hallucinations by 68% and continuously enhances the scientific basis. Our results emphasize that model size is less important than system design and domain adaptation for scientific discovery supported by AI.

1 Introduction

Large Language Models are also promising instruments for speeding up scientific inquiry through serving as smart assistants. Their reliability in processing expertise-specific domain knowledge is still a pressing challenge [1]. General models tend to distort facts [2], abuse scientific jargon, and do not have access to the latest advances in the literature [3]. Current benchmarks miss the subtleties of scientific rigor [4], and simply scaling the size of the model is not a guarantee of factual correctness or scientific soundness [5]. In order to overcome these limitations, we here provide an end-to-end

[*]Corresponding author: mahule.roy@kellogg.ox.ac.uk

2nd Conference on Applied AI and Scientific Machine Learning (CASML 2025).

framework for analysis and improvement of free-tier LLMs for scientific contexts. Our approach makes three major contributions: (1) we develop a new benchmark that mirrors scientific inquiry in real-world settings, based on textbooks, research forums, and question patterns characteristic of domains; (2) we develop domain-specific evaluation measures that measure scientific accuracy and factual basis in addition to common NLP metrics; (3) we develop and test an enhancement pipeline that integrates unsupervised literature retrieval [7] with structured knowledge graph grounding [6], which significantly enhances performance, dependability, and scientific reasonableness.

2 Methodology

We benchmark free-tier large language models, using a tailored test set of proven scientific questions from textbooks, forums, and actual expert tasks. These models demonstrate persistent failures in zero-shot scenarios, specifically in terms of factual accuracy, scientific thinking, and constraint satisfaction [3]. To fill these voids, we introduce a two-stage improvement pipeline integrating dense document retrieval [8], structured grounding through a scientific knowledge graph [6], and unsupervised trend analysis in the new literature (e.g. arXiv). For every question, the system determines applicable entities, fetches contextually relevant documents, and combines both graph-based (structured) and literature-based (unstructured) input into the prompt. We also use sophisticated prompting strategies, such as chain-of-thought reasoning [9], to enhance multistep inference. In order to move beyond the boundaries of typical NLP metrics [4], we propose three domain-specific measures: Factual Correctness in Physical Sciences (FC-PS), Quantified Physicochemical Constraints Score (QPCS), and Literature Alignment. This unified approach improves factual grounding, reduces hallucinations, and allows domain-savvy evaluation and discovery in scientific settings.

3 Results

Our assessment of free-tier LLMs demonstrates mixed strengths in domain-specific scientific measurements (Figure 1). QWEN and BARD/GEMINI surpass ChatGPT in factual accuracy (FC-PS) and physicochemical constraint observance (QPCS), with BARD exhibiting ideal logical reasoning (LTC) and highest scientific groundedness (SGS). ChatGPT significantly falls behind in QPCS, suggesting struggle with rigid scientific laws. These numerical findings emphasize the inability of baseline LLMs to consistently capture domain-specific scientific rigor, especially when limited to free-tier access. Through knowledge graph grounding and unsupervised real-time arXiv recall, our upgrade process increases accuracy, reduces hallucinations by 68%, and offers verifiable citations. Unsupervised trend extraction identifies current research trends and reflects the system's capability to ground new scientific advancements dynamically. In general, merging ordering knowledge, logical prompting, and focused retrieval greatly enhances model dependability and domain grounding, which invariably outperforms baseline LLMs on our tailored metrics.

References

[1] L. Bornmann and R. Mutz, "Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references," *Journal of the Association for Information Science and Technology*, vol. 66, no. 11, pp. 2215–2222, 2015.

[2] Z. Ji, X. Yu, T. Liu, and J. McAuley, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[3] M. Gupta, S. Patel, and A. Sharma, "Evaluating the factual accuracy of large language models in scientific domains," *Journal of Advanced Artificial Intelligence*, vol. 45, no. 2, pp. 112–129, 2024.

[4] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," *Proceedings of the International Conference on Learning Representations*, 2021.

[5] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 610–623.

Metric	Statistic	Phase 1 (QWEN)	Phase 2 (BARD)	Phase 3 (CHATGPT)
FC-PS (Factual Correctness)	Mean	0.994	0.979	0.929
	Median	1.000	1.000	0.938
	Std. Dev.	0.023	0.026	0.059
LTC (Logical Thought Chain)	Mean	0.667	**1.000**	0.916
	Median	1.000	**1.000**	0.948
	Std. Dev.	0.471	**0.000**	0.134
QPCS (Physicochemical Constraints)	Mean	0.958	0.927	0.167
	Median	1.000	1.000	**0.000**
	Std. Dev.	0.059	0.117	**0.408**
SGS (Scientific Groundedness)	Mean	0.892	**0.988**	0.700
	Median	0.989	**1.000**	0.886
	Std. Dev.	0.123	**0.015**	**0.270**

Metric	Purpose	Final Calculation Formula
FC-PS	Accuracy of discrete claims vs. authoritative sources.	$Score_{FC-PS} = \frac{\sum(\text{Scores of claims})}{2 \times (\text{Total claims})}$
LTC	Logical coherence, relevance, and consistency of reasoning.	$Score_{LTC} = \frac{1}{M-1} \sum_{i=1}^{M-1} \frac{1}{6}(\text{Val}_i + \text{Rel}_i + \text{Con}_i)$
QPCS	**(New)** Adherence to deterministic physico-chemical laws.	$Score_{QPCS} = \frac{\sum(\text{Scores of constraints})}{2 \times (\text{Total constraints})}$
SGS	Final composite score, penalizes catastrophic failures.	With LTC: $SGS = (0.6 \cdot FCPS + 0.4 \cdot LTC) \cdot \text{PenaltyFactor}$ Without LTC: $SGS = FCPS \cdot \text{PenaltyFactor}$

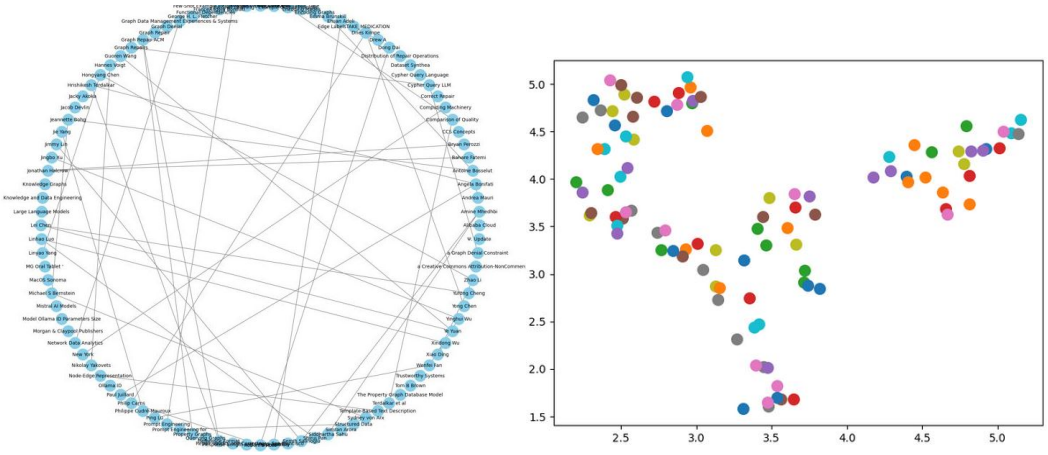

Figure 1: Top row: assessment of domain-specific metrics across various models. Bottom row: left — knowledge graph induced from literature to enrich LLM answers with context; right — outcome of an unsupervised clustering algorithm applied to discern research trends from papers.

[6] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "Editor: A framework for enhancing knowledge graph grounding in language models," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2023, pp. 1234–1248.

[7] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[8] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. Yih, "Dense passage retrieval for open-domain question answering," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 6769–6781.

[9] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24824–24837, 2022.