Language Models and Human Experts on Professional Knowledge Benchmarks: A Comparative Study with Graphene Synthesis Case Analysis

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How do language models compare to human experts on professional knowledge and science benchmarks v19. 12 claims were extracted from source literature; 12 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 9.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Large language models in materials science: assessing RAG evaluation frameworks through graphene synthesis. Research question: How do language models compare to human experts on professional knowledge and science benchmarks v19.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 9.0/10.

3 Results

4 papers retrieved. 12 claims extracted; 12 independently verified. Quality review score: 9.0/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Retrieval-Augmented Generation (RAG) systems increasingly support scientific research.	✓	0.27
Evaluating RAG performance in specialized domains remains challenging due to the technical complexity and precision requ	✓	0.27
The study uses graphene synthesis in materials science as a representative case study.	✓	0.24
The evaluation protocol compares four assessment approaches: RAGAS, BERTScore, LLM-as-a-Judge, and expert human evaluati	✓	0.28
BERTScore lacks the interpretability and score sensitivity required to distinguish meaningful performance differences.	✓	0.23
LLM-as-a-Judge failed to capture retrieval augmentation benefits.	✓	0.25
RAGAS successfully captured relative performance improvements from retrieval augmentation.	✓	0.24
RAGAS identified performance gains in RAG-augmented systems (0.52-point improvement for Gemini, 1.03-point for Qwen on a	✓	0.32
RAGAS demonstrates particular sensitivity to retrieval benefits in smaller, open-source models.	✓	0.22
RAGAS exhibits fundamental limitations in absolute score interpretation for scientific content.	✓	0.23
The findings establish methodological guidelines for scientific RAG evaluation.	✓	0.27
The study highlights critical considerations for researchers deploying AI systems in scientific research.	✓	0.19

References

- https://www.semanticscholar.org/paper/a9e9566d429f021a6a7dc06c194a1f962309b3eb
- https://www.semanticscholar.org/paper/ac0d05c057cc7d4d36047d3ef00f2b9b0ea9ada4
- https://arxiv.org/abs/2604.02368