Prompting Strategies for Maximizing Language Model Accuracy on Graduate-Level Science Questions

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What prompting strategies maximize language model accuracy on graduate-level science questions v19. 17 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: MSQA: Benchmarking LLMs on Graduate-Level Materials Science Reasoning and Knowledge. Research question: What prompting strategies maximize language model accuracy on graduate-level science questions v19.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.0/10.

3 Results

14 papers retrieved. 17 claims extracted; 0 independently verified. Quality review score: 4.0/10.

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
Candidate answers were generated using gpt-4o, gemini-2.0-pro, and deepseek-v3.	×	0.01
Initial LLM-generated answers frequently included ambiguous references such as 'the K0 Samples and the SCAs Units'.	×	0.02
Refining prompts by explicitly discouraging the use of definite articles significantly enhanced answer clarity.	×	0.02
The refined prompts resulted in responses explicitly referencing chemical entities such as hexamethyldisilane and copper	×	0.02
The MSQA dataset evaluation uses two distinct prompting strategies: direct generation and retrieval-augmented generation	×	0.06
In the retrieval-augmented setting, BM25 is used to retrieve the top five most relevant paragraphs to serve as additiona	×	0.03
Black-box LLMs are evaluated exclusively under the direct generation setting.	×	0.07
Evaluation of long-answer responses is conducted through GPT-4o acting as an LLM judge.	×	0.02
In the evaluation metrics, responses categorized as 'correct' or 'mostly correct' are both counted as correct.	×	0.01
For binary-answer evaluations, accuracy is determined by exact keyword matching for responses containing either 'YES' or	×	0.05
The black-box models evaluated include Claude-3.7-Sonnet, Gemini-2.0-Flash, and Grok-3.	×	0.06
The open-source models evaluated include Llama-3-8B, Phi-4-mini, Qwen-2.5-7B, and Deepseek-R1-distilled-Llama-3.	×	0.03
When TPEC aggregates are larger than 2 μm, they are loosely packed, leading to increased fluorescence intensity and a hy	×	0.00
The structural change in TPEC aggregates stabilizes when the space size reduces to 2 μm.	×	0.01
When the liquid phase shrinks to submicrometer dimensions (below 0.5 μm), aggregates are compressed into a more compact	×	0.00
The densification of aggregates in submicrometer dimensions is driven by enhanced solubility in the confined phase, whic 4	×	0.01
Polyhedral oligomeric silsesquioxane (POSS) acts as a nucleator for MgCl2 aggregation in Ziegler-Natta catalysts.	×	0.00

References

- https://arxiv.org/abs/2405.13012
- http://arxiv.org/abs/2401.14043v3
- http://arxiv.org/abs/2505.23982v1