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

Assignee Research

June 5, 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 v8. 12 claims were extracted from source literature; 8 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Large language models encode clinical knowledge. Research question: What prompting strategies maximize language model accuracy on graduate-level science questions v8.

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 7.3/10.

3 Results

14 papers retrieved. 12 claims extracted; 8 independently verified. Quality review score: 7.3/10.

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
MultiMedQA is a benchmark combining six existing medical question answering datasets and a new dataset called HealthSear	✓	0.22
HealthSearchQA is a new dataset consisting of medical questions searched online.	✓	0.16
The proposed human evaluation framework assesses model answers along axes including factuality, comprehension, reasoning	✓	0.24
PaLM is a 540-billion parameter large language model.	✓	0.22
Flan-PaLM is the instruction-tuned variant of PaLM.	✓	0.15
Flan-PaLM achieved state-of-the-art accuracy on the MedQA, MedMCQA, PubMedQA, and MMLU clinical topics datasets within M	✓	0.25
Flan-PaLM achieved 67.6% accuracy on the MedQA dataset.	×	0.15
Flan-PaLM's performance on MedQA surpassed the prior state of the art by more than 17%.	✓	0.16
Human evaluation revealed key gaps in the performance of Flan-PaLM.	×	0.14
Instruction prompt tuning is a parameter-efficient approach for aligning LLMs to new domains using a few exemplars.	✓	0.28
Med-PaLM is the model resulting from applying instruction prompt tuning.	×	0.13
Med-PaLM performs inferior to clinicians.	×	0.14

References

- https://doi.org/10.1145/3560815

- https://doi.org/10.1038/s41586-023-06291-2
- https://doi.org/10.1016/s0140-6736(18)32203-7