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

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What prompting strategies maximize language model accuracy on graduate-level science questions v13. 12 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. Research question: What prompting strategies maximize language model accuracy on graduate-level science questions v13.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

16 papers retrieved. 12 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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
VisualBERT performs the best on average (61.87%) among the VQA baseline models.	×	0.02
Patch-TRM outperforms VisualBERT in natural science (NAT) and language science (LAN).	×	0.04
Patch-TRM performs better in higher-grade questions (67.50% vs. 59.92%).	×	0.02
VisualBERT outperforms Patch-TRM by a large margin (+22.39%) in the subject of social science (SOC).	×	0.03
UniedQA without any supervised ne-tuning (zero-shot) cannot beat any VQA baseline model.	×	0.08
UniedQABASE reports an accuracy of 70.12% on average when ne-tuned with the answer labels in SCIENCEQA.	×	0.03
UniedQABASE (CoT) brings additional improvements of +3.21% (QCM$\rightarrow$AE) and +3.99% (QCM$\rightarrow$ALE).	×	0.03
GPT-3 reaches almost the best performance in the zero-shot setting without any ne-tuning.	×	0.08
The VQA baselines are trained for a maximum number of 50 epochs with a learning rate of 5e-5.	×	0.03
UniedQA is ne-tuned for 50k iterations and evaluated every 1k iteration.	×	0.01
The training process for UniedQA is stopped following the early stopping strategy with a patience period of three evalu	×	0.02
GPT-3 uses the text-davinci-002 engine, which is the most capable model version suggested in the ofcial documentation.	×	0.03

References

- http://arxiv.org/abs/2209.09513v2
- http://arxiv.org/abs/2401.14043v3
- http://arxiv.org/abs/1912.02145v1