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 v10. 13 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Prompt Engineering Strategies for LLM-based Qualitative Coding of Psychological Safety in Software Engineering Communities: A Controlled Empirical Study. Research question: What prompting strategies maximize language model accuracy on graduate-level science questions v10.

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 3.8/10.

3 Results

16 papers retrieved. 13 claims extracted; 0 independently verified. Quality review score: 3.8/10.

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
Cohen's κ (Agreement) is interpreted using the Landis–Koch scale: $<$ 0.20 Slight, 0.20–0.40 Fair, 0.40–0.60 Moderate, 0.6	×	0.02
Intra-Model Stability is measured as the SD of κ across ten independent runs.	×	0.13
Levene's test compares variance between models at each configuration.	×	0.04
Bias Ratio is the ratio of LLM-predicted count for a given category to its human-coded Gold Standard count.	×	0.03
A Bias Ratio of 1.00 indicates no bias.	×	0.02
Values $>$ 1.20 flag over-prediction.	×	0.03
Values $<$ 0.80 flag under-prediction.	×	0.03
Per-class F1 scores are complemented by Wilcoxon signed-rank test (P01 vs. P02, n= 10 pairs, with Cohen's d as effect si	×	0.02
Under P01, κ values ranged from 0.332 (DeepSeek-Chat) to 0.403 (Gemini 2.5 Flash), all within the Fair range.	×	0.11
Under P02, Claude Haiku and Gemini 2.5 Flash crossed into the Moderate range (κ= 0.426 and 0.437 respectively), while De	×	0.13
Across all models, 'Disagreeing with Suggestions or Ideas' achieves the highest F1 scores (0.58–0.70).	×	0.02
Minority categories such as 'Sharing Negative Feedback' (F1 = 0.21–0.30) and 'Admitting Mistakes' (F1 $\approx$0.33–0.40) are sy	×	0.04
Claude Haiku showed a statistically significant improvement from P01 to P02 ($\Delta$$\kappa$= +0.034, p= 0.0).	×	0.07

References

- http://arxiv.org/abs/2505.14347v2
- http://arxiv.org/abs/2605.07422v1
- http://arxiv.org/abs/2510.17892v1