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

Assignee Research

June 5, 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 v7. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.6/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: ChemPro: A Progressive Chemistry Benchmark for Large Language Models. Research question: What prompting strategies maximize language model accuracy on graduate-level science questions v7.

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 4.6/10.

3 Results

16 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 4.6/10.

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
ChemPro consists of 4,100 questions sourced from standardized materials including competitive exams (National Testing Ag	×	0.04
ChemPro spans Inorganic, Organic, Physical and Bio-chemistry.	×	0.12
ChemPro uses educational provenance (NCERT and JEE Mains) to enforce difficulty ordering while spanning high-school chem	×	0.09
ChemPro is a progressive benchmark designed to evaluate LLM chemistry proficiency via a curriculum-aligned progression o	×	0.13
ChemPro experiments span 45 different models, exploring proprietary and open-source variants with varying model sizes.	×	0.09
ChemPro results are reported per tier and subfield.	×	0.02
ARC (Clark et al. 2018) provides elementary to high-school science questions but lacks chemistry-specific depth.	×	0.06
SciBench (Wang et al. 2023) evaluates college-level scientific problem-solving but focuses on undergraduate-to-graduate	×	0.04
JEEBench (Arora, Singh, and Mausam 2023) evaluates high-school problems but lacks systematic difficulty progression with	×	0.05
ChemPro is the only benchmark among the listed ones that is progressive and specifically designed for chemistry.	×	0.09

References

- http://arxiv.org/abs/2410.03595v1

- http://arxiv.org/abs/2604.14214v1
- http://arxiv.org/abs/2602.03108v4