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

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: What prompting strategies maximize language model accuracy on graduate-level science questions v18. 22 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: Exploring Advanced Large Language Models with LLMsuite. Research question: What prompting strategies maximize language model accuracy on graduate-level science questions v18.

2 Methodology

Systematic literature search across multiple databases yielded 12 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

12 papers retrieved. 22 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
BitNet b1.58 achieves up to 2.71 times faster inference and 8.9 times higher throughput than FP16 models of the same siz	×	0.01
BitNet b1.58 is 71.4 times more energy-efficient in matrix multiplication operations compared to FP16 models.	×	0.03
BitNet b1.58 matches or exceeds the performance of FP16 models in terms of perplexity and end-task accuracy, particularl	×	0.04
SuperGLUE is an improved version of GLUE with more challenging tasks.	×	0.04
MMLU is a massive multitask language understanding benchmark.	×	0.06
BIG-Bench is a benchmark for testing LLMs on diverse tasks.	×	0.04
HELM is a holistic evaluation of language models.	×	0.10
Gemini evaluates capabilities of LLMs in medicine.	×	0.04
CoLA is a corpus of linguistic acceptability.	×	0.02
SST2 is the Stanford Sentiment Treebank, used for binary sentiment classification.	×	0.02
MRPC is the Microsoft Research Paraphrase Corpus.	×	0.03
STS is the Semantic Textual Similarity benchmark.	×	0.03
QQP is the Quora Question Pairs dataset.	×	0.01
MNLI is the Multi-Genre Natural Language Inference dataset.	×	0.07
QNLI is the Question Natural Language Inference dataset.	×	0.05
RTE is the Recognizing Textual Entailment dataset.	×	0.01
WNLI is the Winograd NLI dataset.	×	0.01
CoT is the Chain-of-Thought Reasoning dataset.	×	0.02
Muffin is a dataset for fine-tuning LLMs on diverse tasks.	×	0.10
Natural Instructions v2 is a collection of natural language instructions.	×	0.05
MGSM is a dataset for measuring generalization and systematicity in models.	×	0.02
BBH is a dataset for testing generalization beyond Big-Bench.	×	0.01

References

- http://arxiv.org/abs/2304.05302v3
- http://arxiv.org/abs/2406.10833v3
- http://arxiv.org/abs/2407.12036v2