Chain-of-Thought Prompting Effects on LLM Reasoning Accuracy in GSM8K

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How do chain-of-thought prompting strategies impact LLM reasoning accuracy on the GSM8K benchmark compared to standard prompting. 9 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: Instance-adaptive Zero-shot Chain-of-Thought Prompting. Research question: How do chain-of-thought prompting strategies impact LLM reasoning accuracy on the GSM8K benchmark compared to standard prompting?.

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. 9 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
Models tested include LLaMA-3-8B-Instruct, LLaMA-3-70B-Instruct, LLaMA-2-13B-Chat, and Qwen-14B-Chat.	×	0.04
Threshold values for IAP-ss are computed for distinct LLMs on different datasets using Eq 4.	×	0.02
Experiments are run on an 8x NVIDIA A100 GPU server.	×	0.01
Baselines include Answer majority vote (AMV), OPPR, and Self-Discover.	×	0.01
Tasks include GSM8K, SVAMP, CommonsenseQA, MMLU, Causal Judgement, and Tracking Shuffled Objects.	×	0.07
Evaluation metric used is Accuracy for all tasks.	×	0.04
Zero-shot CoT Prompts include 9 specific prompts labeled #1 to #9.	×	0.13
IAP is implemented using 9 prompts as candidates.	×	0.02
Table (p4) and Table (p9) present benchmark results for different models and methods.	×	0.03

References

- http://arxiv.org/abs/2505.03786v1
- http://arxiv.org/abs/2409.20441v3
- http://arxiv.org/abs/2601.03559v2