Chain-of-Thought Prompting and Performance in Long-Horizon LLM Reasoning Tasks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: To what extent does chain-of-thought prompting mitigate performance degradation in long-horizon reasoning tasks for LLMs evaluated on the BigBench suite. 7 claims were extracted from source literature; 7 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. Research question: To what extent does chain-of-thought prompting mitigate performance degradation in long-horizon reasoning tasks for LLMs evaluated on the BigBench suite?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.8/10.

3 Results

15 papers retrieved. 7 claims extracted; 7 independently verified. Quality review score: 8.8/10.

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
Large Language Models (LLMs) can achieve strong performance on many tasks by producing step-by-step reasoning before giv	✓	0.36
CoT explanations can systematically misrepresent the true reason for a model's prediction.	✓	0.28
CoT explanations can be heavily influenced by adding biasing features to model inputs.	✓	0.28
When models are biased toward incorrect answers, they frequently generate CoT explanations rationalizing those answers.	✓	0.25
Accuracy drops by as much as 36% on a suite of 13 tasks from BIG-Bench Hard when testing with GPT-3.5 from OpenAI and Cl	✓	0.26
On a social-bias task, model explanations justify giving answers in line with stereotypes without mentioning the influen	✓	0.34
CoT explanations can be plausible yet misleading, which risks increasing our trust in LLMs without guaranteeing their sa	✓	0.30

References

- https://doi.org/10.48550/arxiv.2305.04388
- https://doi.org/10.48550/arxiv.2305.14314
- https://doi.org/10.48550/arxiv.2301.12726