Perplexity and Downstream Reasoning Performance in Language Models

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: What is the relationship between language model perplexity and downstream reasoning task performance v13. 12 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.1/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Understanding Reasoning in Chain-of-Thought from the Hopfieldian View. Research question: What is the relationship between language model perplexity and downstream reasoning task performance v13.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.1/10.

3 Results

11 papers retrieved. 12 claims extracted; 0 independently verified. Quality review score: 3.1/10.

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
The study evaluates models on six datasets: GSM8K and SVAMP for Arithmetic Reasoning, StrategyQA and CommonsenseQA for C	×	0.06
The Random Letter dataset is constructed from the Last Letter dataset.	×	0.00
The experiments employ Llama-2-7B-Chat, Llama-3-8B-Instruct, Llama-2-13B-Chat, and Llama-2-70B-Chat models.	×	0.05
The study compares three baselines: Base (question only), CoTZ (zero-shot Chain-of-Thought), and CoTF (few-shot Chain-of	×	0.06
On the GSM8K dataset using Llama-2-7B-Chat in the few-shot setting, the CoTF baseline achieved an accuracy of 4.62%.	×	0.03
On the GSM8K dataset using Llama-2-7B-Chat in the few-shot setting, the RoTF method achieved an accuracy of 25.55%.	×	0.03
On the Coin Flip dataset using Llama-3-8B-Instruct in the few-shot setting, the CoTF baseline achieved an accuracy of 96	×	0.03
On the Coin Flip dataset using Llama-3-8B-Instruct in the few-shot setting, the RoTF method achieved an accuracy of 70.3	×	0.04
In the zero-shot setting on the SVAMP dataset with Llama-2-7B-Chat, the RoTZ method achieved an accuracy of 54.33%.	×	0.04
The original CoT performs unstably across different tasks, sometimes performing lower than the Base method.	×	0.03
In the zero-shot scenario on the CSQA dataset, CoT variants performed lower than the Base method.	×	0.03
Compared to CoT prompting, RoT achieves more consistent accuracy improvements across a variety of tasks according to Tab	×	0.07

References

- http://arxiv.org/abs/2503.09567v5
- http://arxiv.org/abs/2410.03595v1
- http://arxiv.org/abs/2503.15113v1