Chain-of-Thought Extended Thinking Benchmark Accuracy Across GSM8K, LogiQA, and BIG-Bench Hard

Assignee Research

June 3, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: Chain-of-thought extended thinking benchmark accuracy improvement survey. 13 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.4/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: CROP: Token-Efficient Reasoning in Large Language Models via Regularized Prompt Optimization. Research question: Chain-of-thought extended thinking benchmark accuracy improvement survey.

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 4.4/10.

3 Results

16 papers retrieved. 13 claims extracted; 1 independently verified. Quality review score: 4.4/10.

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
CROP compresses output token consumption by up to 80.6% without degrading exact-match accuracy.	×	0.07
On the GSM8K dataset, CROP reduced output token consumption by 74.8% while retaining comparable accuracy for Gemini 2.0.	×	0.07
On the Object Counting dataset, CROP reduced output token consumption by 80.6% while retaining comparable accuracy for G	×	0.07
On the LogiQA dataset, CROP reduced output token consumption by 66.5% while retaining comparable accuracy for Gemini 2.0	×	0.07
Evaluating an isolated example (batch size of 1) causes the meta-optimizer to frequently overfit to the length penalty,	×	0.02
Scaling the batch size to 128 stabilizes the optimization process.	×	0.05
CROP introduces an output-biased textual regularizer directly into the prompt optimization loop.	×	0.10
CROP treats output token count as a first-class optimization constraint alongside task accuracy.	×	0.09
CROP computes a continuous length-penalty gradient that functions independently of the task loss.	×	0.03
CROP was evaluated on GSM8K, LogiQA, and BIG-Bench Hard benchmarks.	✓	0.15
In the GSM8K qualitative comparison, Direct Prompting yielded an incorrect answer.	×	0.03
In the GSM8K qualitative comparison, Standard Chain-of-Thought required 128 tokens to produce a correct answer.	×	0.06
In the GSM8K qualitative comparison, CROP required 44 tokens to produce a correct answer.	×	0.03

References

- http://arxiv.org/abs/2503.11495v1
- http://arxiv.org/abs/2604.14214v1

- http://arxiv.org/abs/2604.14140v1