State-of-the-Art Reasoning Performance in RL-Fine-Tuned Large Language Models

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: What are the state-of-the-art large language model results on reasoning benchmarks published recently v20. 16 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Large Language Models Reasoning Abilities Under Non-Ideal Conditions After RL-Fine-Tuning. Research question: What are the state-of-the-art large language model results on reasoning benchmarks published recently v20.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.2/10.

3 Results

4 papers retrieved. 16 claims extracted; 3 independently verified. Quality review score: 5.2/10.

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 paper investigates the reasoning performance of LMs fine-tuned with RL in non-ideal scenarios.	✓	0.18
The paper proposes a new research direction inspired by brain science: examining the reasoning abilities of RL-fine-tune	✓	0.27
The paper demonstrates that RL-fine-tuned LMs exhibit significant performance degradation under non-ideal scenarios.	✓	0.16
The paper designs effective remediation strategies tailored to each scenario by manipulating the format reward and the e	×	0.02
The paper publicly releases high-quality, novel evaluation datasets designed to assess LM performance under noisy condit	×	0.03
The paper uses multiple choice questions that involve multiple possibilities and require a conclusive answer to address	×	0.10
The data samples used to evaluate RQ1 for LLMs are drawn from CommonsenseQA and Ceval-exam.	×	0.04
CommonsenseQA has 2000 training samples, 500 validation samples, and 1000 test samples.	×	0.01
Ceval-exam has 700 training samples, 246 validation samples, and 400 test samples.	×	0.01
MathVision has 700 training samples, 200 validation samples, and 500 test samples.	×	0.01
WeThink has 3000 training samples, 1000 validation samples, and 2000 test samples.	×	0.01
SciVQA has 2000 training samples, 400 validation samples, and 1000 test samples.	×	0.01
Math12k has 2500 training samples, 500 validation samples, 388 test samples in TestA, 745 test samples in TestB, 388 tes	×	0.01
MathReasoning has 3000 training samples, 500 validation samples, 468 test samples in TestA, 1000 test samples in TestB,	×	0.01
Mathverse has 280 training samples, 90 validation samples, 115 test samples in TestA, 114 test samples in TestB, 115 tes	×	0.01
MathVision has 1400 training samples, 400 validation samples, 336 test samples in TestA, 480 test samples in TestB, 336	×	0.01

References

- http://arxiv.org/abs/2407.08029v1
- http://arxiv.org/abs/2508.04848v1
- http://arxiv.org/abs/2309.02144v1