Reinforcement Learning from Human Feedback Enhances Language Model Mathematical Reasoning

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How does reinforcement learning from human feedback improve language model mathematical reasoning v12. 8 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Learning to Plan with Natural Language. Research question: How does reinforcement learning from human feedback improve language model mathematical reasoning v12.

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 3.0/10.

3 Results

4 papers retrieved. 8 claims extracted; 0 independently verified. Quality review score: 3.0/10.

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 method involves reasoning types: mathematical reasoning, causal reasoning, logical reasoning, symbolic reasoning, an	×	0.05
The method uses a compression prompt to reduce repeated solutions, trivial information, and token cost.	×	0.04
The method updates the plan by appending the new plan update to the end of the current task plan.	×	0.12
The method uses a compression prompt to generate a compressed and shorter plan if the task plan is too long.	×	0.13
ChatGPT + Learning-to-Plan (zero-shot, CoT) achieves an average score of 66.6 across 10 tasks.	×	0.05
ChatGPT + Learning-to-Plan (zero-shot, CoT) achieves an average score of 29.8 across 7 tasks.	×	0.06
ChatGPT + Learning-to-Plan (zero-shot, CoT) achieves an average score of 66.0 across 7 tasks.	×	0.06
GPT-4-32k achieves scores of 57.8, 64.6, 81.3, and 85.0 in Algebra, Causal Judgment, Logical Reasoning (LSAT), and Last	×	0.01

References

- http://arxiv.org/abs/2407.17482v2
- http://arxiv.org/abs/2308.04332v1
- http://arxiv.org/abs/2304.10464v4