Scalable Preference Model Pretraining for Enhanced LLM Mathematical Reasoning

Assignee Research

June 7, 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 v20. 10 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: CodePMP: Scalable Preference Model Pretraining for Large Language Model Reasoning. Research question: How does reinforcement learning from human feedback improve language model mathematical reasoning 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 4.2/10.

3 Results

4 papers retrieved. 10 claims extracted; 1 independently verified. Quality review score: 4.2/10.

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
CodePMP achieves higher RM accuracy on both 1.5B and 7B models across mathematical and logical reasoning tasks.	×	0.09
CodePMP achieves higher BoN accuracy on both mathematical and logical reasoning tasks for all model sizes.	×	0.08
CodePMP models maintain accuracy up to N=256, while non-CodePMP models exhibit significant accuracy degradation at highe	×	0.04
CodePMP is a highly scalable method.	×	0.06
CodePMP uses deepseek-coder-6.7b-instruct for generating code preference pairs.	×	0.10
CodePMP uses MetaMath-Mistral-7B as the generator for BoN evaluation.	×	0.01
CodePMP is evaluated on GSM8K and MATH for mathematical reasoning, and ReClor and LogiQA2.0 for logical reasoning.	×	0.14
CodePMP uses multiple-choice accuracy (equivalent to Best-of-4) for logical reasoning tasks.	×	0.05
CodePMP exhibits strong generalization, yielding significant improvements across different reasoning tasks.	×	0.08
CodePMP is a scalable preference model pretraining pipeline that enhances LLM reasoning abilities using synthesized pref	✓	0.34

References

- http://arxiv.org/abs/2410.02884v2
- http://arxiv.org/abs/2406.10858v2
- http://arxiv.org/abs/2410.02229v2