Scaling Preference Datasets in CodePMP and Diminishing Returns in Reward Model Alignment

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: Does scaling the size of the preference dataset in CodePMP yield diminishing returns in reward model alignment scores on the HH-RLHF benchmark. 16 claims were extracted from source literature; 2 were 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: CodePMP: Scalable Preference Model Pretraining for Large Language Model Reasoning. Research question: Does scaling the size of the preference dataset in CodePMP yield diminishing returns in reward model alignment scores on the HH-RLHF benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 14 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

14 papers retrieved. 16 claims extracted; 2 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
CodePMP achieves higher RM accuracy on both 1.5B and 7B models across mathematical and logical reasoning tasks.	×	0.11
CodePMP achieves higher BoN accuracy on both mathematical and logical reasoning tasks for all model sizes.	×	0.09
CodePMP models maintain accuracy up to N=256, while non-CodePMP models exhibit significant accuracy degradation at highe	×	0.03
CodePMP-initialized models outperform baselines across various N values, showing superior ranking capabilities.	×	0.03
CodePMP exhibits strong generalization, yielding significant improvements across different reasoning tasks.	×	0.10
CodePMP is a highly scalable method.	×	0.07
CodePMP enhances the model's ability to differentiate correct from incorrect reasoning.	×	0.04
CodePMP uses synthesized preference pairs derived from high-quality, publicly available source code.	✓	0.22
CodePMP uses deepseek-coder-6.7b-instruct for data construction.	×	0.03
CodePMP uses MetaMath-Mistral-7B as the generator for BoN evaluation.	×	0.03
CodePMP evaluates on GSM8K and MATH for mathematical reasoning, and ReClor and LogiQA2.0 for logical reasoning.	✓	0.17
CodePMP uses multiple-choice accuracy (equivalent to Best-of-4) for logical reasoning tasks.	×	0.07
CodePMP achieves RM accuracy of 0.4154, 0.4804, 0.5351, 0.3665, 0.2751 for 1.5B model and 0.5839, 0.4972, 0.5022, 0.5240	×	0.03
CodePMP achieves BoN accuracy of 0.6126, 0.9050, 0.4364, 0.3698, 0.6041 for 1.5B model and 0.7668, 0.9413, 0.5373, 0.490	×	0.02
CodePMP achieves BoN accuracy of 0.6841 for 1.5B model and 0.6912 for 7B model.	×	0.03
CodePMP achieves multiple-choice accuracy of 0.758 for 1.5B model and 0.7619 for 7B model.	×	0.03

References

- http://arxiv.org/abs/2402.18571v3
- http://arxiv.org/abs/2508.04149v2
- http://arxiv.org/abs/2410.02229v2