Explicit vs Implicit Reward Modeling Effects on Alignment Quality and Training Stability

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: How do different reward modeling approaches (implicit vs explicit rewards) influence the final alignment quality and training stability on the SQuTR benchmark. Current large language models (LLMs) often struggle to produce accurate responses on the first attempt for complex reasoning tasks like code generation. Prior research tackles this challenge by generating multiple candidate solutions and validating them with LLM-generated unit. 9 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Dynamic Scaling of Unit Tests for Code Reward Modeling. Research question: How do different reward modeling approaches (implicit vs explicit rewards) influence the final alignment quality and training stability on the SQuTR benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.0/10.

3 Results

11 papers retrieved. 9 claims extracted; 3 independently verified. Quality review score: 6.0/10.

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
CodeRM-8B improves the performance of Llama3-8B by 18.43% on HumanEval Plus.	✓	0.17
CodeRM-8B enhances the performance of Llama3-70B by 4.95% on HumanEval Plus.	×	0.08
CodeRM-8B enhances the performance of GPT-4o-mini by 3.42% on HumanEval Plus.	×	0.14
Dynamic unit test scaling brings up to approximately 0.5% performance gain on MBPP Plus at a fixed computational cost.	×	0.11
There is a positive correlation between the number of unit tests and the quality of the code reward signal.	✓	0.31
Scaling unit tests is more effective for harder problems.	✓	0.19
Gemma-2-27B-it achieves comparable performance to Llama3.1-70B when scaled to 100 unit tests per question.	×	0.08
The unit test-based majority voting framework follows a standard best-of-N strategy.	×	0.06
The optimal candidate solution is selected based on majority voting.	×	0.02

References

- http://arxiv.org/abs/2409.13948v3
- http://arxiv.org/abs/2501.01054v1
- http://arxiv.org/abs/2310.05910v2