Socratic Feedback Loops in Reward Models for Enhanced HumanEval Pass@k Scores

Assignee Research

June 11, 2026

Abstract

Large Language Models (LLMs) can generate code from natural language, but their performance is highly sensitive to prompt formulation. We propose a reinforcement-learning-based framework that models prompt refinement as a sequential decision-making problem. A Proximal Policy Optimization (PPO) agent iteratively improves prompts using a hybrid action space that combines direct generation, genetic lexical mutation and semantic rewriting, guided by shaped rewards derived from unit-test feedback. We evaluate the framework on MBPP+, HumanEval+, and APPS using CodeT5+, CodeLLaMA, and DeepSeek-Coder

1 Introduction

This paper examines: Prompt Optimization for LLM Code Generation via Reinforcement Learning. Research question: How does integrating Socratic feedback loops into reward models affect pass@k scores on the HumanEval benchmark compared to standard binary reward signals?.

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 7.2/10.

3 Results

4 papers retrieved. 15 claims extracted; 13 independently verified. Quality review score: 7.2/10.

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 proposed framework models prompt refinement as a sequential decision-making problem using reinforcement learning.	✓	0.28
The framework utilizes a Proximal Policy Optimization (PPO) agent.	✓	0.16
The agent uses a hybrid action space combining direct generation, genetic lexical mutation, and semantic rewriting.	✓	0.24
The agent is guided by shaped rewards derived from unit-test feedback.	✓	0.23
The framework was evaluated on the MBPP+, HumanEval+, and APPS benchmarks.	×	0.13
CodeT5+, CodeLLaMA, and DeepSeek-Coder were used as frozen code generators in the evaluation.	✓	0.17
On the 500-task MBPP+ test set, the PPO agent achieved a strict Pass@1 score of 57.58% with CodeT5+.	✓	0.20
On the 500-task MBPP+ test set, the PPO agent achieved a strict Pass@1 score of 64.80% with CodeLLaMA.	✓	0.20
On the 500-task MBPP+ test set, the PPO agent achieved a strict Pass@1 score of 85.50% with DeepSeek-Coder.	✓	0.21
On the MBPP+ test set, the proposed method outperformed EPiC, Reflexion, and Random-Hybrid baselines.	×	0.13
On the MBPP+ test set, the PPO agent achieved a Soft-Pass@1 score of 67.90% with CodeT5+.	✓	0.17
On the MBPP+ test set, the PPO agent achieved a Soft-Pass@1 score of 73.10% with CodeLLaMA.	✓	0.17
On the MBPP+ test set, the PPO agent achieved a Soft-Pass@1 score of 88.20% with DeepSeek-Coder.	✓	0.18
Similar performance improvements were observed on HumanEval+ and APPS benchmarks across all tested backbone models.	✓	0.19
Reinforcement learning with shaped test-driven rewards improves functional correctness in LLM-based code generation.	✓	0.38

References

- https://doi.org/10.5281/zenodo.18394393
- https://doi.org/10.48550/arxiv.2504.09037
- https://openalex.org/W7162044111