Multi-Objective Reward Alignment vs. Scalar-Reward RLHF in Code Generation Latency and Throughput

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does multi-objective reward alignment compare to scalar-reward RLHF in terms of inference latency and throughput when evaluated on the DS-1000 code generation benchmark. Q-shaping is an extension of Q-value initialization and serves as an alternative to reward shaping for incorporating domain knowledge to accelerate agent training, thereby improving sample efficiency by directly shaping Q-values. This approach is both general and robust across. 10 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: From Reward Shaping to Q-Shaping: Achieving Unbiased Learning with LLM-Guided Knowledge. Research question: How does multi-objective reward alignment compare to scalar-reward RLHF in terms of inference latency and throughput when evaluated on the DS-1000 code generation benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.3/10.

3 Results

15 papers retrieved. 10 claims extracted; 1 independently verified. Quality review score: 4.3/10.

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
Q-shaping improves performance by 16.87% compared to the best baseline across 20 tasks.	×	0.15
Q-shaping improves performance by 55.39% compared to TD3.	×	0.04
Q-shaping outperforms LLM-based reward shaping methods by 253.80% in peak performance improvement.	✓	0.24
LLM-TD3 improved by 38.68% in the door-close task compared to the best baseline.	×	0.05
LLM-TD3 improved by 406.04% in the drawer-open task compared to the best baseline.	×	0.05
LLM-TD3 improved by 389.77% in the window-close task compared to the best baseline.	×	0.05
LLM-TD3 improved by 180.70% in the sweep-into task compared to the best baseline.	×	0.05
o1-Preview achieved 100% template adherence in generating heuristic functions.	×	0.01
GPT-4o achieved 100% correct Q-values in generating heuristic functions.	×	0.06
Gemini achieved only 44% average correctness in generating heuristic functions.	×	0.01

References

- http://arxiv.org/abs/2406.12845v1
- http://arxiv.org/abs/2410.01458v1

- http://arxiv.org/abs/2506.08062v2