Q-Shaping vs. Reward Shaping Performance in HH-RLHF Harmlessness Across Instructional Datasets

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: Does Q-shaping maintain consistent performance gains over reward shaping in terms of HH-RLHF harmlessness scores when applied to LLMs trained with different instructional datasets (e.g., Alpaca vs.. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: The Art of Efficient Reasoning: Data, Reward, and Optimization. Research question: Does Q-shaping maintain consistent performance gains over reward shaping in terms of HH-RLHF harmlessness scores when applied to LLMs trained with different instructional datasets (e.g., Alpaca vs. Vicuna)?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.8/10.

3 Results

12 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 2.8/10.

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
Training exclusively on hard prompts results in catastrophic failure, with policy entropy spiking drastically and rollou	×	0.02
Training on easier prompts yields the most stable trajectory, with low and stable policy entropy, and rollout length ada	×	0.02
Despite training on easy prompts, the performance on relatively tough tasks (e.g., AIME'25) is comparable to (or even sl	×	0.02
Increasing the rollout number N yields observable benefits, significantly speeding up the Length Adaptation phase and le	×	0.07
Larger N leads to a faster recovery of reasoning capabilities and higher asymptotic Mean@8 in mathematical benchmarks.	×	0.03
The performance gap due to varying N is task-dependent, with more significant benefits observed in mathematical benchmar	×	0.02
Training on relatively easier prompts provides a denser positive reward signal, which is essential for stable reasoning	×	0.06
More rollouts contribute to better performance but also bring heavier training costs.	×	0.03
The learned length bias can be generalized across domains, i.e., training on mathematical prompts works well on the code	×	0.08
The effectiveness of different strategies is budget-dependent, exhibiting distinct or even contradictory behaviors.	×	0.02

References

- http://arxiv.org/abs/2502.21321v2
- http://arxiv.org/abs/2602.20945v3
- http://arxiv.org/abs/2410.01458v1