Potential-Based Reward Shaping Scaling Effects on Token Efficiency and Harmlessness in HH-RLHF

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: Does the effectiveness of potential-based reward shaping in reducing inference token count without sacrificing harmlessness scores on the HH-RLHF dataset degrade or improve as model size scales from. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.3/10. 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: Does the effectiveness of potential-based reward shaping in reducing inference token count without sacrificing harmlessness scores on the HH-RLHF dataset degrade or improve as model size scales from 7B to 70B compared to state-based methods?.

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 6.3/10.

3 Results

4 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 6.3/10.

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.

References

- http://arxiv.org/abs/2502.01307v1
- http://arxiv.org/abs/2405.19595v1
- http://arxiv.org/abs/2410.01458v1