Reward Shaping Strategies and Cross-Domain Alignment Generalization in Reinforcement Learning

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How do different reward shaping strategies (e.g., intrinsic vs. extrinsic) affect the generalization of alignment across diverse domains (e.g., reasoning, multimodal tasks), as measured by. 12 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.1/10. 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: How do different reward shaping strategies (e.g., intrinsic vs. extrinsic) affect the generalization of alignment across diverse domains (e.g., reasoning, multimodal tasks), as measured by cross-domain HH-RLHF helpfulness and harmlessness scores?.

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 4.1/10.

3 Results

15 papers retrieved. 12 claims extracted; 0 independently verified. Quality review score: 4.1/10.

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 (pass rate $\leq$ 0.5 over N=8 rollouts) results in catastrophic failure, characterized	×	0.03
Training on hard prompts causes downstream performance metrics, such as Mean@8 on AMC and Olympiad Bench, to degrade sig	×	0.04
Training on easy prompts (pass rate > 0.5) yields a stable training trajectory with low and stable policy entropy.	×	0.02
Despite training exclusively on easy prompts, performance on tough tasks like AIME'25 is comparable to or slightly excee	×	0.03
Increasing the rollout number N (tested at 8, 12, 16, 24) significantly speeds up the Length Adaptation phase, causing t	×	0.03
While larger rollout numbers (N) lead to faster convergence in length adaptation, all tested settings (N=8 to N=24) conv	×	0.04
Larger rollout numbers (N) lead to a more robust Reasoning Refinement stage, resulting in faster recovery of reasoning c	×	0.07
The performance benefit of increasing rollout number N is task-dependent, with different effects observed on LiveCodeBen	×	0.03
The learned length bias from training on mathematical prompts generalizes to code tasks.	×	0.10
The effectiveness of different reasoning strategies is dependent on the token budget, exhibiting distinct or contradicto	×	0.11
Using the 'Art' method reduces output token length by percentages ranging from 14.5% to 47.3% across Qwen3 models from 0	×	0.09
Using the 'Art' method improves AIME'25 Mean@8 scores by up to 11.2 points across Qwen3 models ranging from 0.6B to 30B	×	0.09

References

- `http://arxiv.org/abs/1804.06459v2`

- http://arxiv.org/abs/2602.20945v3
- http://arxiv.org/abs/2410.01729v1