Potential-Based vs. State-Based Rewards in Zero-Shot Cross-Domain NLP Transfer

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What is the comparative effect of potential-based vs. state-based reward functions on the zero-shot cross-domain transfer capabilities of 7B and 70B models evaluated on the SLM-Bench NLP tasks. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.0/10. 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: What is the comparative effect of potential-based vs. state-based reward functions on the zero-shot cross-domain transfer capabilities of 7B and 70B models evaluated on the SLM-Bench NLP tasks outside CommonsenseQA?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.0/10.

3 Results

16 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 2.0/10.

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.	×	0.02
Training on the easier counterpart yields the most stable trajectory.	×	0.01
The performance on relatively tough tasks (e.g., AIME'25) is comparable to (or even slightly exceeding) training on the	×	0.02
Increasing N yields observable benefits that significantly speed up the Length Adaptation phase.	×	0.05
With a larger N, it is easier to discover short and correct trajectories.	×	0.07
Larger N leads to a more robust Reasoning Refinement stage.	×	0.07
The model recovers its reasoning capabilities faster and achieves a higher asymptotic Mean@8 with larger N.	×	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.	×	0.04
The learned length bias can be generalized across domains, i.e., training on mathematical prompts works well on the code	×	0.09

References

- http://arxiv.org/abs/2509.16679v1
- http://arxiv.org/abs/2602.20945v3
- http://arxiv.org/abs/2604.25872v1