Scaling Model Size and Reward Functions in CommonsenseQA Robustness Evaluation

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How does model size scaling (from 7B to 70B) influence the robustness of CommonsenseQA performance when comparing potential-based and state-based reward functions across different hardware. 6 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.0/10. 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: How does model size scaling (from 7B to 70B) influence the robustness of CommonsenseQA performance when comparing potential-based and state-based reward functions across different hardware configurations?.

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 5.0/10.

3 Results

16 papers retrieved. 6 claims extracted; 0 independently verified. Quality review score: 5.0/10.

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 with drastic policy entropy spikes and premature ro	×	0.03
Training on easy prompts yields stable training dynamics with low and stable policy entropy.	×	0.02
Despite training on easy prompts, performance on tough tasks (e.g., AIME'25) is comparable to training on the full datas	×	0.04
Increasing rollout number N speeds up the Length Adaptation phase and leads to faster decay in rollout length.	×	0.05
Larger rollout numbers lead to more robust Reasoning Refinement and higher asymptotic Mean@8 on mathematical benchmarks.	×	0.03
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/2312.17661v1
- http://arxiv.org/abs/2602.20945v3
- http://arxiv.org/abs/2402.11651v2