Oracle-RLAIF CIDEr Gains Over SFT vs Reinforcement Learning Baselines on MSVD

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does the CIDEr score improvement of Oracle-RLAIF over SFT compare to other reinforcement learning methods (e.g., PPO, DQN) on the MSVD benchmark across different model sizes. In post-training for reasoning Large Language Models (LLMs), the current state of practice trains LLMs in two independent stages: Supervised Fine-Tuning (SFT) and Reinforcement Learning with Verifiable Rewards (RLVR, shortened as "RL" below). In this work, we challenge whether. 6 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Quagmires in SFT-RL Post-Training: When High SFT Scores Mislead and What to Use Instead. Research question: How does the CIDEr score improvement of Oracle-RLAIF over SFT compare to other reinforcement learning methods (e.g., PPO, DQN) on the MSVD benchmark across different model sizes?.

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 4.5/10.

3 Results

12 papers retrieved. 6 claims extracted; 0 independently verified. Quality review score: 4.5/10.

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
Model performance is measured as Pass@1 averaged over 64 repetitions across 7 math benchmarks.	×	0.07
Evaluations are conducted with pipelines based on vllm and HuggingFace's math-verify for efficient inference.	×	0.02
The evaluation uses decoding temperature t=1.0 and the standard reasoning prompt.	×	0.02
Prediction accuracy based on SFT Pass@1 baseline achieves an R2 coefficient of 0.57 on Llama3-8B-Instruct model.	×	0.06
Spearman's Rank Correlation for prediction based on SFT Pass@Large k (k=64) achieves 0.94 for Llama3-8B-Instruct and 0.9	×	0.14
Coefficient of determination (R2) for prediction based on SFT Generalization Loss achieves 0.88 for Llama3-8B-Instruct a	×	0.12

References

- http://arxiv.org/abs/2602.07464v1
- http://arxiv.org/abs/2510.01624v1
- http://arxiv.org/abs/2510.02561v1