Robustness of RLAIF-Trained Non-Autoregressive Multimodal Models vs. SFT Baselines

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does the robustness of RLAIF-trained non-autoregressive multimodal models compare to SFT baselines in terms of accuracy on adversarial or low-quality video inputs, as measured by COCO-Caption or. We give simpler, sparser, and faster algorithms for differentially private fine-tuning of large-scale pre-trained language models, which achieve the state-of-the-art privacy versus utility tradeoffs on many standard NLP tasks. We propose a meta-framework for this problem. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Differentially Private Fine-tuning of Language Models. Research question: How does the robustness of RLAIF-trained non-autoregressive multimodal models compare to SFT baselines in terms of accuracy on adversarial or low-quality video inputs, as measured by COCO-Caption or SPICE 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 6.5/10.

3 Results

15 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 6.5/10.

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.

References

- http://arxiv.org/abs/2007.08428v4
- http://arxiv.org/abs/2110.06500v2
- http://arxiv.org/abs/1904.10076v3