Oracle-RLAIF and RLHF Efficiency Trade-offs on SQuTR with Noisy Spoken Queries

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 3 peer-reviewed papers addressing the following research question: What is the efficiency trade-off between Oracle-RLAIF and RLHF in terms of inference latency and memory usage when processing noisy spoken queries on the SQuTR benchmark. While large-scale unsupervised language models (LMs) learn broad world knowledge and some reasoning skills, achieving precise control of their behavior is difficult due to the completely unsupervised nature of their training. Existing methods for gaining such steerability. 13 claims were extracted from source literature; 12 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Direct Preference Optimization: Your Language Model is Secretly a Reward Model. Research question: What is the efficiency trade-off between Oracle-RLAIF and RLHF in terms of inference latency and memory usage when processing noisy spoken queries on the SQuTR benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 3 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.2/10.

3 Results

3 papers retrieved. 13 claims extracted; 12 independently verified. Quality review score: 8.2/10.

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
Large-scale unsupervised language models learn broad world knowledge and some reasoning skills.	✓	0.26
Achieving precise control of language model behavior is difficult due to the completely unsupervised nature of their tra	✓	0.26
Existing methods for steerability collect human labels of the relative quality of model generations and fine-tune the un	✓	0.43
RLHF is a complex and often unstable procedure.	✓	0.16
RLHF involves first fitting a reward model that reflects human preferences, and then fine-tuning the large unsupervised	✓	0.40
The paper introduces a new parameterization of the reward model in RLHF that enables extraction of the corresponding opt	✓	0.29
The new parameterization allows solving the standard RLHF problem with only a simple classification loss.	✓	0.17
The resulting algorithm is called Direct Preference Optimization (DPO).	✓	0.21
DPO eliminates the need for sampling from the LM during fine-tuning.	✓	0.19
DPO eliminates the need for performing significant hyperparameter tuning.	×	0.15
Experiments show that DPO can fine-tune LMs to align with human preferences as well as or better than existing methods.	✓	0.34
Fine-tuning with DPO exceeds PPO-based RLHF in the ability to control sentiment of generations.	✓	0.30
Fine-tuning with DPO matches or improves response quality compared to existing methods.	✓	0.20

References

- https://doi.org/10.48550/arxiv.2312.14925
- https://doi.org/10.48550/arxiv.2305.18290

- https://doi.org/10.48550/arxiv.2404.18930