Reward-Free vs. Reward-Based Alignment for LLM Robustness Against Adversarial Prompts

Assignee Research

June 5, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: What is the impact of reward-free alignment methods like DPO versus reward-based RLHF on the robustness of LLMs against adversarial prompts in safety evaluation datasets. 10 claims were extracted from source literature; 1 was 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: Inference-Time Safety For Code LLMs Via Retrieval-Augmented Revision. Research question: What is the impact of reward-free alignment methods like DPO versus reward-based RLHF on the robustness of LLMs against adversarial prompts in safety evaluation datasets?.

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. 10 claims extracted; 1 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
SOSECURE is evaluated on three datasets: SALLM, LLMSecEval, and a subset of LMSys-Chat-1M.	×	0.02
SALLM contains 100 security-oriented prompts, with 74 samples used for evaluation after filtering.	×	0.03
LLMSecEval consists of 150 prompts, with 49 Python and 40 C samples used for evaluation after filtering.	×	0.02
The LMSys subset contains 240 Python samples used for analysis after a two-stage filtration process.	×	0.02
GPT-4 is used as the underlying code generation model across all experiments.	×	0.10
SOSECURE is compared against three baselines: Prompt-only, Revision-only, and GPT-4+CWE.	×	0.04
Security outcomes are assessed using CodeQL and Bandit.	×	0.02
SOSECURE is designed to improve the security of LLM-generated code by incorporating community-authored security knowledg	✓	0.16
SOSECURE is model-agnostic and requires no retraining or fine-tuning.	×	0.08
SOSECURE performs three steps: retrieval of security-relevant community discussions, construction of a revision prompt,	×	0.15

References

- http://arxiv.org/abs/2509.09055v1

- http://arxiv.org/abs/2306.11066v2
- http://arxiv.org/abs/2603.01494v1