Reverse-KL Regularization in RLHF Enhances Robustness of Vision-Language Models on VQA-Adv

Assignee Research

June 1, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: How does reverse-KL regularization in RLHF impact the robustness of vision-language models against adversarial perturbations on the VQA-Adv benchmark. In the last few years, the deep learning (DL) computing paradigm has been deemed the Gold Standard in the machine learning (ML) community. Moreover, it has gradually become the most widely used computational approach in the field of ML, thus achieving outstanding results on. 4 claims were extracted from source literature; 2 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: Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Research question: How does reverse-KL regularization in RLHF impact the robustness of vision-language models against adversarial perturbations on the VQA-Adv benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 11 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

11 papers retrieved. 4 claims extracted; 2 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.

5 Extracted Claims

Claim	Verified	Confidence
Deep learning has achieved outstanding results on several complex cognitive tasks, matching or even beating human perfor	✓	0.21
Deep learning has outperformed well-known machine learning techniques in domains such as cybersecurity, natural language	✓	0.28
Existing reviews on Deep Learning only tackle one aspect of the field, leading to an overall lack of knowledge.	×	0.11
The proposed review uses a holistic approach to provide a comprehensive survey of Deep Learning aspects and recent enhan	×	0.14

References

- https://doi.org/10.48550/arxiv.2001.06937
- https://doi.org/10.48550/arxiv.2307.16851
- https://doi.org/10.1186/s40537-021-00444-8