Retrieval-Augmented Generation Effects on Vision-Language Model Robustness Against Multimodal Jailbreaks

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does Retrieval-Augmented Generation impact the robustness of vision-language models against multimodal jailbreak attacks on benchmarks like MM-SafetyBench compared to non-RAG baselines. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: BlueSuffix: Reinforced Blue Teaming for Vision-Language Models Against Jailbreak Attacks. Research question: How does Retrieval-Augmented Generation impact the robustness of vision-language models against multimodal jailbreak attacks on benchmarks like MM-SafetyBench compared to non-RAG baselines?.

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 5.8/10.

3 Results

15 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 5.8/10.

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/2505.21556v1
- http://arxiv.org/abs/2410.20971v2
- http://arxiv.org/abs/2405.18770v6