Alignment Techniques and Robustness of Qwen14B on GSM8K-V Under Adversarial Noise

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: To what extent do alignment techniques (e.g., instruction tuning vs. reinforcement learning from human feedback) improve the robustness of Qwen14B on GSM8K-V when exposed to adversarial visual noise,. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: On the Adversarial Robustness of Vision Transformers. Research question: To what extent do alignment techniques (e.g., instruction tuning vs. reinforcement learning from human feedback) improve the robustness of Qwen14B on GSM8K-V when exposed to adversarial visual noise, and how does this compare to models like LLaVA-1.5 and InstructBLIP?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.5/10.

3 Results

13 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 2.5/10.

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/2509.25160v1
- http://arxiv.org/abs/2402.11690v1
- http://arxiv.org/abs/2103.15670v3