VISTA Reveals Robustness Gaps in Llama-3.1-8B Across Adversarial Dialogue Benchmarks

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: Does VISTA reveal significant robustness discrepancies in open-weight models like Llama-3.1-8B when evaluated on adversarial multi-turn dialogues versus standard conversational benchmarks. 12 claims were extracted from source literature; 8 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: MTMCS-Bench: Evaluating Contextual Safety of Multimodal Large Language Models in Multi-Turn Dialogues. Research question: Does VISTA reveal significant robustness discrepancies in open-weight models like Llama-3.1-8B when evaluated on adversarial multi-turn dialogues versus standard conversational benchmarks?.

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 7.5/10.

3 Results

12 papers retrieved. 12 claims extracted; 8 independently verified. Quality review score: 7.5/10.

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
MTMCS-Bench contains over 30 thousand multimodal (image+text) and unimodal (text-only) samples.	✓	0.26
MTMCS-Bench offers paired safe and unsafe dialogues with structured evaluation.	✓	0.24
MTMCS-Bench evaluates contextual safety in MLLMs under two complementary settings: escalation-based risk and context-swi	✓	0.35
MTMCS-Bench comprises 752 base images and 2,256 variants, with 12,032 dialogues and 18,048 questions, totaling 30,080 sa	×	0.05
Each sample in MTMCS-Bench is provided in both multimodal and unimodal formats.	×	0.07
MTMCS-Bench introduces a comprehensive evaluation framework that measures contextual safety from three complementary per	✓	0.17
MTMCS-Bench combines multi-turn multimodal contextual safety with paired safe/unsafe variants over the same scenes, text	✓	0.21
MTMCS-Bench has 30,080 samples.	×	0.08
MTMCS-Bench includes multi-turn, image variants, unimodal counterpart, and MCQ/TF evaluation.	×	0.09
MTMCS-Bench evaluates eight open-source and seven proprietary MLLMs.	✓	0.20
Models tend to either miss gradual risks or over-refuse benign dialogues.	✓	0.21
Five current guardrails mitigate some failures but do not fully resolve multi-turn contextual risks.	✓	0.29

References

- http://arxiv.org/abs/2504.07887v2
- http://arxiv.org/abs/2601.06757v1
- http://arxiv.org/abs/2310.13650v1