Qwen2.5 and Prior Versions Safety Alignment on Multimodal Adversarial Benchmarks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: What is the comparative safety alignment performance of Qwen2.5 models versus prior versions on adversarial benchmarks like RedBench or WildQA, measured by safety score variance across different. 15 claims were extracted from source literature; 9 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.0/10. 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: What is the comparative safety alignment performance of Qwen2.5 models versus prior versions on adversarial benchmarks like RedBench or WildQA, measured by safety score variance across different dialogue turns?.

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 7.0/10.

3 Results

11 papers retrieved. 15 claims extracted; 9 independently verified. Quality review score: 7.0/10.

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.	✓	0.23
MTMCS-Bench evaluates contextual safety in MLLMs under two complementary settings: escalation-based risk and context-swi	✓	0.33
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.	×	0.06
MTMCS-Bench introduces a comprehensive evaluation framework that measures contextual safety from three complementary per	✓	0.15
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.	×	0.06
MTMCS-Bench includes multi-turn, image variants, unimodal counterpart, and MCQ/TF evaluation.	×	0.10
MTMCS-Bench evaluates models on Type A (Overall) and Type B (Overall) with metrics including MCQ (%), TF (%), Safety Awa	×	0.08
MTMCS-Bench evaluates eight open-source and seven proprietary MLLMs.	✓	0.19
MTMCS-Bench evaluates five current guardrails.	×	0.11
MTMCS-Bench observes persistent trade-offs between contextual safety and utility in MLLMs.	✓	0.22
MTMCS-Bench finds that models tend to either miss gradual risks or over-refuse benign dialogues.	✓	0.21
MTMCS-Bench finds that current guardrails mitigate some failures but do not fully resolve multi-turn contextual risks.	✓	0.29

References

- http://arxiv.org/abs/2602.04796v1
- http://arxiv.org/abs/2601.06757v1
- http://arxiv.org/abs/2512.17083v3