Robustness of Multimodal Model Alignment on Adversarial and Out-of-Distribution Uni-MMMU Samples

Assignee Research

June 3, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How robust are current alignment techniques in multimodal models when evaluated on adversarial or out-of-distribution samples from Uni-MMMU's science and engineering disciplines. 7 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Uni-MMMU: A Massive Multi-discipline Multimodal Unified Benchmark. Research question: How robust are current alignment techniques in multimodal models when evaluated on adversarial or out-of-distribution samples from Uni-MMMU's science and engineering disciplines?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

4 papers retrieved. 7 claims extracted; 1 independently verified. Quality review score: 4.5/10.

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
Uni-MMMU is a benchmark that evaluates unified models in an integrated manner, highlighting the synergy between their un	✓	0.18
Uni-MMMU includes four key dimensions: multimodal understanding (MMU), generation and editing (Gen&Edit), multi-turn eva	×	0.08
Mazes in Uni-MMMU are procedurally generated using DFS carving to ensure a loop-free structure and BFS verification to g	×	0.03
Only mazes with shortest path lengths between 2 and 10 are retained in Uni-MMMU.	×	0.04
The Sliding Puzzle task in Uni-MMMU evaluates optimal state-space search and visual execution.	×	0.07
Models in the Sliding Puzzle task are given the initial and goal states of a 3 $\times$ 3 8-puzzle rendered in a fixed 9-color	×	0.02
Puzzles in the Sliding Puzzle task are generated by applying random moves to the solved state, then using BFS to verify	×	0.02

References

- http://arxiv.org/abs/1909.08072v2
- http://arxiv.org/abs/2510.13759v3
- http://arxiv.org/abs/2405.18770v6