Quantized InternLM Scaling Effects on Adversarial Multimodal Stability in LLaVA

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: How does the scaling of quantized InternLM models (7B vs. 13B) influence performance stability in the presence of adversarial multimodal inputs compared to full-precision baselines on the LLaVA. We introduce LLaVA-Critic, the first open-source large multimodal model (LMM) designed as a generalist evaluator to assess performance across a wide range of multimodal tasks. LLaVA-Critic is trained using a high-quality critic instruction-following dataset that incorporates. 11 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: LLaVA-Critic: Learning to Evaluate Multimodal Models. Research question: How does the scaling of quantized InternLM models (7B vs. 13B) influence performance stability in the presence of adversarial multimodal inputs compared to full-precision baselines on the LLaVA benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.0/10.

3 Results

10 papers retrieved. 11 claims extracted; 0 independently verified. Quality review score: 4.0/10.

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
LLaVA-Critic achieves a Pearson-r score of 0.673 on ImageDC, 0.706 on MMVet, 0.580 on Wild-Vision, 0.529 on LLaVA-B, 0.82	×	0.04
LLaVA-Critic-7B (v0.5) achieves a Pearson-r score of 0.737 on ImageDC, 0.718 on MMVet, 0.571 on WildVision, 0.494 on LLa	×	0.04
LLaVA-Critic-7B achieves a Pearson-r score of 0.735 on ImageDC, 0.733 on MMVet, 0.616 on WildVision, 0.494 on LLaVA-B, 0	×	0.04
GPT-4o achieves an accuracy of 0.617 with ties and 0.734 without ties, with a Kendall's τ of 0.819.	×	0.01
GPT-4V achieves an accuracy of 0.620 with ties and 0.733 without ties, with a Kendall's τ of 0.787.	×	0.04
LLaVA-Critic-7B achieves an accuracy of 0.596 with ties and 0.722 without ties, with a Kendall's τ of 0.763.	×	0.05
LLaVA-Critic-72B achieves an accuracy of 0.605 with ties and 0.736 without ties, with a Kendall's τ of 0.779.	×	0.05
GPT-4V* achieves a model score of 0.490, pairwise accuracy with ties of 0.636, and pairwise accuracy without ties of 0.7	×	0.04
GPT-4o achieves a model score of 0.439, pairwise accuracy with ties of 0.577, and pairwise accuracy without ties of 0.7	×	0.01
LLaVA-Critic-7B achieves a model score of 0.314, pairwise accuracy with ties of 0.556, and pairwise accuracy without tie	×	0.05
LLaVA-Critic-72B achieves a model score of 0.287, pairwise accuracy with ties of 0.513, and pairwise accuracy without ti	×	0.04

References

- http://arxiv.org/abs/2407.17856v4
- http://arxiv.org/abs/2404.01331v2
- http://arxiv.org/abs/2410.02712v2