Mixed-Precision Quantization Trade-offs in Multimodal Models: Efficiency vs. Reasoning Accuracy

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: What is the trade-off between inference efficiency (latency/throughput) and reasoning accuracy when applying mixed-precision quantization to multimodal models like InternLM on benchmarks such as MMMU. Reinforcement learning (RL) has become a key technique for enhancing the reasoning abilities of large language models (LLMs), with policy-gradient algorithms dominating the post-training stage because of their efficiency and effectiveness. However, most existing benchmarks. 11 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Large Language Models Reasoning Abilities Under Non-Ideal Conditions After RL-Fine-Tuning. Research question: What is the trade-off between inference efficiency (latency/throughput) and reasoning accuracy when applying mixed-precision quantization to multimodal models like InternLM on benchmarks such as MMMU or SEED-Bench?.

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 5.2/10.

3 Results

10 papers retrieved. 11 claims extracted; 3 independently verified. Quality review score: 5.2/10.

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
The paper investigates the reasoning performance of LMs fine-tuned with RL in non-ideal scenarios through three research	✓	0.26
The baselines used in the experiments include Qwen 2.5-VL-7B-Instruct, Llama 3.1-8B-Instruct, Qwen 3-14B, and Mistral-Sm	×	0.04
The datasets used for Research Question 1 include CommonsenseQA and Ceval-exam.	×	0.01
CommonsenseQA has 2000 training samples, 500 validation samples, and 1000 test samples.	×	0.01
Ceval-exam has 700 training samples, 246 validation samples, and 400 test samples.	×	0.01
The datasets used for Research Questions 2 and 3 include Math12k, MathReasoning, Mathverse, and MathVision.	×	0.02
Math12k has 2500 training samples, 500 validation samples, 388 TestA samples, 745 TestB samples, 388 FineTest samples, a	×	0.00
The paper proposes a new research direction inspired by brain science: examining the reasoning abilities of RL-fine-tune	✓	0.28
The paper demonstrates that RL-fine-tuned LMs exhibit significant performance degradation under non-ideal scenarios.	✓	0.15
The paper designs effective remediation strategies tailored to each scenario by manipulating the format reward and the e	×	0.02
The paper publicly releases high-quality, novel evaluation datasets designed to assess LM performance under noisy condit	×	0.05

References

- http://arxiv.org/abs/2401.16420v1
- http://arxiv.org/abs/2508.04848v1
- http://arxiv.org/abs/2309.15112v5