

Quantized MobileVLM vs Full-Precision VLMs on MME and MM1K Benchmarks

Assignee Research

June 5, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How do quantized MobileVLM models (1.4B and 2.7B) compare to full-precision 3B-13B VLMs on MME and MM1K benchmarks in terms of reasoning accuracy and inference latency. 8 claims were extracted from source literature; 1 was 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: Do VLMs Truly "Read" Candlesticks? A Multi-Scale Benchmark for Visual Stock Price Forecasting. Research question: How do quantized MobileVLM models (1.4B and 2.7B) compare to full-precision 3B-13B VLMs on MME and MM1K benchmarks in terms of reasoning accuracy and inference latency?.

2 Methodology

Systematic literature search across multiple databases yielded 14 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

14 papers retrieved. 8 claims extracted; 1 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
VLMs exhibit strong predictive capability in persistent uptrend or downtrend conditions but weak predictive capability i	✓	0.20
VLMs show significant prediction biases and limited sensitivity to explicitly specified forecast horizons in prompts.	×	0.10
The dataset contains 69744 daily and 123780 weekly candlestick charts for HS300 and S&P500 respectively.	×	0.07
Claude-sonnet-4-5 (thinking) model achieved an accuracy of 51.46% and an F1 score of 0.3496 on the HS300 market.	×	0.01
Gemini2.5-pro (thinking) model achieved an accuracy of 51.37% and an F1 score of 0.5722 on the S&P 500 market.	×	0.02
Gemini2.5-flash (nothinking) model has a mean rank of 0.02576 and a median rank of 0.05374 in the IC (Complete Ranking T	×	0.01
Gemini2.5-pro (thinking) model has a mean rank of 0.03047 and a median rank of 0.07221 in the IC (Complete Ranking Table	×	0.01
XGBoost(30d) model has a mean rank of -0.01030 and a median rank of 0.03596 in the IC (Complete Ranking Table).	×	0.02

References

- <http://arxiv.org/abs/2411.03314v1>
- <http://arxiv.org/abs/2604.12659v1>
- <http://arxiv.org/abs/2306.09265v1>