DeepSeek-V4-Pro Inference Efficiency on MMLU and GSM8K Benchmarks

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does the inference efficiency of DeepSeek-V4-Pro compare to other LLMs on standard reasoning benchmarks like MMLU and GSM8K. Rapid advancements in large language models (LLMs) have increased interest in deploying them on mobile devices for on-device AI applications. Mobile users interact differently with LLMs compared to desktop users, creating unique expectations and data biases. 9 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Mobile-MMLU: A Mobile Intelligence Language Understanding Benchmark. Research question: How does the inference efficiency of DeepSeek-V4-Pro compare to other LLMs on standard reasoning benchmarks like MMLU and GSM8K?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.3/10.

3 Results

13 papers retrieved. 9 claims extracted; 0 independently verified. Quality review score: 2.3/10.

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 Mobile-MMLU benchmark consists entirely of multiple-choice questions.	×	0.13
Llama-3.2-1B-instruct achieves a lower performance on Mobile-MMLU compared to larger models such as Gemma-2-9B-it.	×	0.06
Qwen2.5-3B-Instruct achieves 68.1% accuracy on Mobile-MMLU.	×	0.05
Llama-3.2-3B-Instruct scores 50.2% accuracy on Mobile-MMLU.	×	0.05
The performance spread on MMLU ranges from 45.9% to 71.8%.	×	0.02
The performance spread on MMLU-Pro ranges from 7.5% to 36.5%.	×	0.07
The performance spread on Mobile-MMLU ranges from 34.5% to 75.0%.	×	0.06
Model Llama-3.2 (1B) achieves 46.84% accuracy on Mobile-MMLU.	×	0.05
Phi-3.5-mini-instruct achieves 63.7% accuracy on Mobile-MMLU.	×	0.05

References

- http://arxiv.org/abs/2312.17080v4
- http://arxiv.org/abs/2503.20786v1
- http://arxiv.org/abs/2401.02954v1