How does the PowerInfer approach scale in terms of memory efficiency and throughput when applied to multi-moda

Assignee Research

May 29, 2026

Abstract

The field of efficient Large Language Model (LLM) inference is rapidly evolving, presenting a unique blend of opportunities and challenges. Although the field has expanded and is vibrant, there hasn't been a concise framework that analyzes the various methods of LLM Inference to provide a clear understanding of this domain. Our survey stands out from traditional literature reviews by not only summarizing the current state of research but also by introducing a framework based on roofline model for systematic analysis of LLM inference techniques. This framework identifies the bottlenecks when de

1 Introduction

This paper examines: LLM Inference Unveiled: Survey and Roofline Model Insights. Research question: How does the PowerInfer approach scale in terms of memory efficiency and throughput when applied to multi-modal LLM inference (e.g., LLaVA) on consumer-grade GPUs compared to dense inference methods?.

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 6.3/10.

3 Results

10 papers retrieved. 9 claims extracted; 4 independently verified. Quality review score: 6.3/10.

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
There has not been a concise framework that analyzes the various methods of LLM Inference to provide a clear understandi	✓	0.30
The survey introduces a framework based on the roofline model for systematic analysis of LLM inference techniques.	✓	0.27
The proposed framework identifies bottlenecks when deploying LLMs on hardware devices.	✓	0.18
The framework provides an explanation for why LLMs are memory-bound.	×	0.11
The framework quantifies the memory and computation requirements of LLMs.	×	0.07
The survey covers model compression techniques including Knowledge Distillation and Quantization.	×	0.12
The survey covers algorithm improvements including Early Exit and Mixture-of-Expert.	×	0.15
The survey covers hardware and system-level enhancements.	×	0.12
The survey analyzes inference methods using the roofline model to determine their impact on memory access and computatio	✓	0.19

References

- https://doi.org/10.48550/arxiv.2402.16363
- https://doi.org/10.48550/arxiv.2312.15234
- https://doi.org/10.48550/arxiv.2401.08092