PowerInfer Sparsity Optimization and Inference Latency in LLaMA Scaling Across GPU Configurations

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does PowerInfer's neuron activation sparsity optimization affect inference latency when scaling from LLaMA-33B to LLaMA-70B across different consumer GPU memory configurations. This paper introduces PowerInfer, a high-speed Large Language Model (LLM) inference engine on a personal computer (PC) equipped with a single consumer-grade GPU. The key principle underlying the design of PowerInfer is exploiting the high locality inherent in LLM inference. 4 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: PowerInfer: Fast Large Language Model Serving with a Consumer-grade GPU. Research question: How does PowerInfer's neuron activation sparsity optimization affect inference latency when scaling from LLaMA-33B to LLaMA-70B across different consumer GPU memory configurations?.

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 4.2/10.

3 Results

13 papers retrieved. 4 claims extracted; 0 independently verified. Quality review score: 4.2/10.

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
PowerInfer delivers an average generation speed of 13.20 tokens/s for quantized models and 8.32 tokens/s for non-quantiz	×	0.10
PowerInfer's performance is up to 8.00$\times$ better than llama.cpp for quantized models and 11.69$\times$ for non-quantized models.	×	0.06
The inference speed on an NVIDIA RTX 4090 GPU is only 18% slower compared to the performance on a top-tier A100 GPU.	×	0.14
Only a limited number of neurons are activated during each inference iteration in LLMs, due to activation sparsity.	×	0.11

References

- http://arxiv.org/abs/2312.12456v2
- http://arxiv.org/abs/2511.02213v1
- http://arxiv.org/abs/2005.14187v1