What is the impact of PowerInfer's neuron activation sparsity on throughput and latency when running LLaMA-33B versus LLaMA-70B on consumer-grade GPUs

Assignee Research

May 29, 2026

Abstract

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, characterized by a power-law distribution in neuron activation. This distribution indicates that a small subset of neurons, termed hot neurons, are consistently activated across inputs, while the majority, cold neurons, vary based on specific inputs. PowerInfer exploits such an insight to design a GPU-CPU hybrid inference

1 Introduction

This paper examines: PowerInfer: Fast Large Language Model Serving with a Consumer-grade GPU. Research question: What is the impact of PowerInfer's neuron activation sparsity on throughput and latency when running LLaMA-33B versus LLaMA-70B on consumer-grade GPUs?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

12 papers retrieved. 13 claims extracted; 1 independently verified. Quality review score: 4.5/10.

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, deployed on a PC equipped with a single NVIDIA RTX 4090 GPU, delivers an average generation speed of 13.20 t	✓	0.15
PowerInfer exhibits up to 8.00$\times$ and 11.69$\times$ improvements for quantized and non-quantized models, respectively, compared t	×	0.05
The inference speed achieved on an NVIDIA RTX 4090 GPU (priced at approximately $2,000) is only 18% slower compared to t	×	0.11
PowerInfer's code has been open sourced completely.	×	0.03
LLM inference is an autoregressive model that generates each token based on previous ones.	×	0.07
The LLM architecture includes multiple Transformer layers, each comprising a self-attention and an MLP (Multi-Layer Perc	×	0.07
The self-attention block generates embedding vectors by capturing the relationships among input tokens.	×	0.01
The MLP block applies non-linear transformations via fully connected layers and activation functions to refine the input	×	0.06
Activation functions (like ReLU and SiLU) act as gates to selectively retain or discard values in a vector, influencing	×	0.03
State-of-the-art systems like llama.cpp distribute layers between CPU and GPU memories, leveraging both for inference, t	×	0.09
Current hardware architectures are designed with a memory hierarchy optimized for data locality.	×	0.04
Each LLM inference iteration requires accessing the entire set of model parameters whose total size is too large for a s	×	0.07
Recent works have identified activation sparsity in LLM inference.	×	0.08

References

- `http://arxiv.org/abs/2210.06313v2`

- http://arxiv.org/abs/2601.11743v1
- http://arxiv.org/abs/2312.12456v2