SOVEREIGN: What is the inference latency and throughput trade-off for LLaMA models of varying sizes (7B to 70B) when eval

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

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 e

1 Introduction

Analysis of: PowerInfer: Fast Large Language Model Serving with a Consumer-grade GPU. Research goal: What is the inference latency and throughput trade-off for LLaMA models of varying sizes (7B to 70B) when evaluated on the HumanEval and GSM8K benchmarks, and how does this efficiency compare to state-of-the-art sparse or quantized models like GPTQ or SparseGPT?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

9 papers retrieved. 0 claims extracted, 0 verified. Tribunal: 6.5/10 $\rightarrow$ REVISE (revision_round=1). Policy: ESCALATE_TO_OWNER.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

References

- https://doi.org/10.48550/arxiv.2312.12456
- https://doi.org/10.1145/3694715.3695964
- https://doi.org/10.48550/arxiv.2307.10169