PowerInfer Hot Neuron Threshold Effects on LLaMA-33B and LLaMA-70B Inference Trade-offs

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How does the PowerInfer hot neuron activation threshold parameter impact inference latency and accuracy trade-offs for LLaMA-33B and LLaMA-70B on the HumanEval code generation benchmark. Deploying local AI models, such as Large Language Models (LLMs), to edge devices can substantially enhance devices' independent capabilities, alleviate the server's burden, and lower the response time. Owing to these tremendous potentials, many big tech companies have released. 12 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.1/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Activation Sparsity Opportunities for Compressing General Large Language Models. Research question: How does the PowerInfer hot neuron activation threshold parameter impact inference latency and accuracy trade-offs for LLaMA-33B and LLaMA-70B on the HumanEval code generation benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.1/10.

3 Results

16 papers retrieved. 12 claims extracted; 0 independently verified. Quality review score: 3.1/10.

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
Activation sparsity can be used to achieve compression benefits in LLMs without altering the activation function.	×	0.09
Our approach omits lower magnitude values to induce sparsity, offering a novel method for LLM compression.	×	0.04
The systematic exploration indicates that 50% extra sparsity in FFN layers can be achieved with negligible accuracy loss	×	0.09
State-of-the-art LLMs such as Falcon and LLaMa use non-ReLU activation functions like GELU and Swish.	×	0.06
Recent LLMs such as Falcon and LLaMa use GELU and Swish as their activation functions.	×	0.03
The ReLUfication technique introduces sparse ReLU-based activations into non-ReLU LLMs.	×	0.04
Directly swapping activation functions often fails to achieve satisfactory sparsity due to unaddressed inherent limitati	×	0.04
Major tech companies are investing in edge LLM solutions.	×	0.06
Edge devices face restrictions in computing and memory resources that limit effective LLM execution.	×	0.07
Existing compression techniques are insufficient for most edge devices to execute LLMs effectively.	×	0.11
New compression opportunities from activation sparsity can benefit general LLMs and are not bound by specific activation	×	0.12
It is possible to safely secure 50% extra sparsity in FFN layers with negligible accuracy loss for state-of-the-art LLMs	×	0.13

References

- http://arxiv.org/abs/2210.06313v2
- http://arxiv.org/abs/2412.12178v2
- http://arxiv.org/abs/2410.12381v3