Dynamic Hot Neuron Threshold Adjustment in PowerInfer for LLaMA-70B Inference Efficiency

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How does the dynamic hot neuron threshold adjustment in PowerInfer compare to fixed threshold methods in terms of inference latency and memory efficiency when applied to LLaMA-70B on MBPP Python. Large Language Models achieve remarkable performance but incur substantial computational costs unsuitable for resource-constrained deployments. This paper presents the first comprehensive task-specific efficiency analysis comparing 16 language models across five diverse NLP. 15 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Task-Specific Efficiency Analysis: When Small Language Models Outperform Large Language Models. Research question: How does the dynamic hot neuron threshold adjustment in PowerInfer compare to fixed threshold methods in terms of inference latency and memory efficiency when applied to LLaMA-70B on MBPP Python function synthesis tasks?.

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 5.0/10.

3 Results

16 papers retrieved. 15 claims extracted; 1 independently verified. Quality review score: 5.0/10.

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
The study evaluates 16 representative open-source language models ranging from 0.5B to 72B parameters.	×	0.09
The IMDB Movie Reviews dataset benchmark consists of 1,000 movie reviews for binary sentiment classification.	×	0.03
The HellaSwag benchmark includes 10,042 examples for commonsense reasoning using log-likelihood scoring.	×	0.03
The ARC-Easy benchmark comprises 2,376 multiple-choice questions regarding elementary scientific knowledge.	×	0.03
The SQuAD 2.0 benchmark contains 11,873 examples for reading comprehension requiring answer generation or unanswerable q	×	0.03
The GSM8K benchmark includes 1,319 grade-school problems for multi-step mathematical reasoning.	×	0.03
The Performance-Efficiency Ratio (PER) metric combines accuracy, throughput, memory usage, and latency.	✓	0.20
The PER metric calculation uses a geometric mean after min-max normalization of dimensions to the [0, 1] range.	×	0.07
On the GSM8K benchmark, the Llama-3.1-8B model achieved an accuracy of 0.8097.	×	0.03
On the GSM8K benchmark, the Qwen2.5-0.5B model achieved a throughput of 7927 tokens/s.	×	0.02
On the GSM8K benchmark, the Llama-3.1-8B model has a latency of 3.74 ms/token.	×	0.04
On the HellaSwag benchmark, the Qwen2.5-72B model achieved an accuracy of 0.64.	×	0.02
On the HellaSwag benchmark, the Llama-3.1-8B model has a latency of 132 ms/sample.	×	0.04
The Qwen2.5-72B model achieved a PER score of 0 on the GSM8K benchmark.	×	0.02
The Qwen2.5-0.5B model achieved a PER score of 0 on the HellaSwag benchmark.	×	0.02

References

- http://arxiv.org/abs/2312.12456v2
- http://arxiv.org/abs/2603.21389v1
- http://arxiv.org/abs/2303.12869v1