What is the correlation between varying hot neuron selection thresholds in PowerInfer and token generation thr

Assignee Research

May 29, 2026

Abstract

Small language models (SLMs), despite their widespread adoption in modern smart devices, have received significantly less academic attention compared to their large language model (LLM) counterparts, which are predominantly deployed in data centers and cloud environments. While researchers continue to improve the capabilities of LLMs in the pursuit of artificial general intelligence, SLM research aims to make machine intelligence more accessible, affordable, and efficient for everyday tasks. Focusing on transformer-based, decoder-only language models with 100M-5B parameters, we survey 70 state

1 Introduction

This paper examines: Small Language Models: Survey, Measurements, and Insights. Research question: What is the correlation between varying hot neuron selection thresholds in PowerInfer and token generation throughput when running LLaMA-70B on the MBPP code generation dataset?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 9.2/10.

3 Results

4 papers retrieved. 11 claims extracted; 11 independently verified. Quality review score: 9.2/10.

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
Small language models (SLMs) have received significantly less academic attention compared to large language models (LLMs	✓	0.32
SLMs are predominantly deployed in modern smart devices.	✓	0.16
LLMs are predominantly deployed in data centers and cloud environments.	✓	0.21
Researchers continue to improve the capabilities of LLMs in the pursuit of artificial general intelligence.	✓	0.28
SLM research aims to make machine intelligence more accessible, affordable, and efficient for everyday tasks.	✓	0.31
The survey focuses on transformer-based, decoder-only language models with 100M-5B parameters.	✓	0.26
The survey analyzes 70 state-of-the-art open-source SLMs.	✓	0.20
The analysis covers technical innovations across three axes: architectures, training datasets, and training algorithms.	✓	0.24
The capabilities of SLMs are evaluated in various domains, including commonsense reasoning, mathematics, in-context lear	✓	0.28
The paper benchmarks the inference latency and memory footprints of SLMs to gain insight into their on-device runtime co	✓	0.22
The paper offers valuable insights to advance research in the field of SLMs through in-depth analysis of benchmarking da	✓	0.22

References

- https://doi.org/10.48550/arxiv.2409.15790
- https://doi.org/10.48550/arxiv.2404.14294
- https://doi.org/10.1145/3695053.3731092