PowerInfer Dynamic Hot Neuron Thresholding for LLaMA-70B Inference Latency Reduction

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: What is the relative inference latency improvement of PowerInfer's dynamic hot neuron threshold adjustment compared to static baselines for LLaMA-70B on the same MBPP Python function synthesis. This investigation aims to study different adaptive fuzzy inference algorithms capable of real-time sequential learning and prediction of time-series data. A brief qualitative description of these algorithms namely meta-cognitive fuzzy inference system (McFIS), sequential. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: A Comparative Study: Adaptive Fuzzy Inference Systems for Energy Prediction in Urban Buildings. Research question: What is the relative inference latency improvement of PowerInfer's dynamic hot neuron threshold adjustment compared to static baselines for LLaMA-70B on the same MBPP Python function synthesis benchmark?.

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 6.2/10.

3 Results

12 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 6.2/10.

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.

References

- http://arxiv.org/abs/2502.17521v2
- http://arxiv.org/abs/2010.15748v4
- http://arxiv.org/abs/1809.08860v1