Cold Neuron Pruning and Reasoning Accuracy in PowerInfer Inference Pipelines

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does neuron activation sparsity correlate with reasoning task accuracy degradation when models are pruned to cold neurons only in PowerInfer's inference pipeline. Activation sparsity offers a compelling route to accelerate large language model (LLM) inference by selectively suppressing hidden activations, yet existing approaches exhibit severe accuracy degradation at high sparsity. We show that this failure stems from representational. 12 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Resting Neurons, Active Insights: Robustifying Activation Sparsity in LLMs via Spontaneity. Research question: How does neuron activation sparsity correlate with reasoning task accuracy degradation when models are pruned to cold neurons only in PowerInfer's inference pipeline?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.0/10.

3 Results

13 papers retrieved. 12 claims extracted; 0 independently verified. Quality review score: 3.0/10.

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
SPON exhibits remarkable representational resilience across various quantization precisions, maintaining stable perplexi	×	0.04
SPON under int4 quantization consistently outperforms the baseline results reported in the full 32-bit setting.	×	0.02
SPON provides a protective effect against network pruning.	×	0.05
SPON helps recover the information density lost during weight-space pruning.	×	0.01
SPON opens new frontiers for extreme model compression.	×	0.05
SPON is inference-neutral.	×	0.06
SPON helps to approximate the full dense model's performance across different sparsity levels and heterogeneous backbone	×	0.06
At moderate sparsity (25%), SPON exhibits negligible perplexity degradation.	×	0.05
As sparsity increases to more aggressive regimes (50% and beyond), purely activation-based methods such as TEAL experien	×	0.04
SPON consistently mitigates this degradation across all evaluated backbones.	×	0.07
SPON outperforms SOTA network pruning methods.	×	0.03
SPON does not rely on using the same dataset for calibration and evaluation.	×	0.03

References

- http://arxiv.org/abs/2512.12744v4
- http://arxiv.org/abs/2307.05639v2
- http://arxiv.org/abs/2412.12178v2