What is the correlation between privacy budget epsilon and inference latency for LLaMA-2 when evaluated on the

Assignee Research

June 10, 2026

Abstract

We find arithmetic ability resides within a limited number of attention heads, with each head specializing in distinct operations. To delve into the reason, we introduce the Comparative Neuron Analysis (CNA) method, which identifies an internal logic chain consisting of four distinct stages from input to prediction: feature enhancing with shallow FFN neurons, feature transferring by shallow attention layers, feature predicting by arithmetic heads, and prediction enhancing among deep FFN neurons. Moreover, we identify the human-interpretable FFN neurons within both feature-enhancing and feature

1 Introduction

This paper examines: Interpreting Arithmetic Mechanism in Large Language Models through Comparative Neuron Analysis. Research question: What is the correlation between privacy budget epsilon and inference latency for LLaMA-2 when evaluated on the SVAMP benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.5/10.

3 Results

15 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 6.5/10.

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/2003.00973v2
- http://arxiv.org/abs/0803.3946v4
- http://arxiv.org/abs/2409.14144v1