Inference Optimization Trade-offs in Llama3-70B and Codestral-34B for Infill Tasks

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: What is the impact of different inference optimization techniques on the latency and accuracy trade-off between Llama3-70B and Codestral-34B for SIMCOPILOT's infill tasks. Deep ensemble learning has been shown to improve accuracy by training multiple neural networks and averaging their outputs. Ensemble learning has also been suggested to defend against membership inference attacks that undermine privacy. 8 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Accuracy-Privacy Trade-off in Deep Ensemble: A Membership Inference Perspective. Research question: What is the impact of different inference optimization techniques on the latency and accuracy trade-off between Llama3-70B and Codestral-34B for SIMCOPILOT's infill tasks?.

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 4.8/10.

3 Results

4 papers retrieved. 8 claims extracted; 1 independently verified. Quality review score: 4.8/10.

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
There is an increase in both accuracy and MI attack effectiveness as we go from a single model to ensembles comprising o	×	0.05
The trade-off between accuracy and privacy is more noticeable for more accurate models trained for a larger number of ep	×	0.08
Starting with a single well-trained model achieving around 70% test accuracy as a baseline (for non-ensemble case), ense	×	0.08
Ensembling can be adopted to improve privacy by intentionally using an ensemble of underfitted models instead of a sing	×	0.08
These two objectives (improving accuracy and improving privacy) are not achieved simultaneously in deep ensembles.	×	0.10
Using deep ensembles to improve accuracy exacerbates its susceptibility to membership inference attacks by making train	✓	0.19
The most widely-used form of ensembling in deep models is deep ensembles.	×	0.09
The most common type of membership inference attack is based on confidence outputs.	×	0.13

References

- http://arxiv.org/abs/2208.13968v1
- http://arxiv.org/abs/2105.05381v4
- http://arxiv.org/abs/2408.15301v2