Trade-offs Between Model Size, Latency, and Accuracy for OpenPangu-7B-MLA Versus Prosody-Exclusive Models on Edge Devices

Assignee Research

June 12, 2026

Abstract

To help MLOps engineers decide which operator to use in which deployment scenario, this study aims to empirically assess the accuracy vs latency trade-off of white-box (training-based) and black-box operators (non-training-based) and their combinations in an Edge AI setup. We perform inference experiments including 3 white-box (i.e., QAT, Pruning, Knowledge Distillation), 2 black-box (i.e., Partition, SPTQ), and their combined operators (i.e., Distilled SPTQ, SPTQ Partition) across 3 tiers (i.e., Mobile, Edge, Cloud) on 4 commonly-used Computer Vision and Natural Language Processing models to

1 Introduction

This paper examines: On the Impact of White-box Deployment Strategies for Edge AI on Latency and Model Performance. Research question: How does the trade-off between model size and latency compare between OpenPangu-7B-MLA and smaller prosody-exclusive models when deployed on edge devices for real-time EchoMind classification, and what is the impact on accuracy under fixed hardware constraints?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.5/10.

3 Results

14 papers retrieved. 7 claims extracted; 7 independently verified. Quality review score: 8.5/10.

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.

5 Extracted Claims

Claim	Verified	Confidence
The study empirically assesses the accuracy vs latency trade-off of white-box and black-box operators and their combinat	✓	0.30
Inference experiments include 3 white-box operators (QAT, Pruning, Knowledge Distillation), 2 black-box operators (Parti	✓	0.36
Experiments are performed across 3 tiers (Mobile, Edge, Cloud) on 4 commonly-used Computer Vision and Natural Language P	✓	0.28
The combination of Distillation and SPTQ operators (DSPTQ) should be preferred over non-hybrid operators when lower late	✓	0.42
Among non-hybrid operators, the Distilled operator is a better alternative in both mobile and edge tiers for lower laten	✓	0.46
Operators involving distillation show lower latency in resource-constrained tiers (Mobile, Edge) compared to operators i	✓	0.39
For textual subject models with low input data size requirements, the Cloud tier is a better alternative for deployment	✓	0.37

References

- http://arxiv.org/abs/2403.17154v4
- http://arxiv.org/abs/2601.22873v1
- http://arxiv.org/abs/2411.00907v3