What is the computational efficiency trade-off between OpenPangu-7B-MLA and prosody-exclusive models when depl

Assignee Research

June 10, 2026

Abstract

Generative Artificial Intelligence (GenAI) applies models and algorithms such as Large Language Model (LLM) and Foundation Model (FM) to generate new data. GenAI, as a promising approach, enables advanced capabilities in various applications, including text generation and image processing. In current practice, GenAI algorithms run mainly on the cloud server, leading to high latency and raising security concerns. Consequently, these challenges encourage the deployment of GenAI algorithms directly on edge devices. However, the large size of such models and their significant computational resourc

1 Introduction

This paper examines: GenAI at the Edge: Comprehensive Survey on Empowering Edge Devices. Research question: What is the computational efficiency trade-off between OpenPangu-7B-MLA and prosody-exclusive models when deployed on edge devices for real-time EchoMind classification, measured by latency and throughput under fixed hardware constraints?.

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 4.0/10.

3 Results

15 papers retrieved. 3 claims extracted; 0 independently verified. Quality review score: 4.0/10.

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
GenAI models have increasingly large model architectures that present significant deployment challenges.	×	0.11
Early attempts to address deployment challenges explored distributed mobile computing systems that could partition model	×	0.06
Research in enabling broader deployment and access of GenAI models has focused on three principal directions.	×	0.05

References

- http://arxiv.org/abs/2601.08844v1
- http://arxiv.org/abs/2504.03656v1
- http://arxiv.org/abs/2502.15816v1