How does the accuracy and inference latency of lightweight BERT models compare to distilled versions of larger

Assignee Research

May 29, 2026

Abstract

The deployment of transformer-based models on resource-constrained edge devices represents a critical challenge in enabling real-time artificial intelligence applications. This comprehensive survey examines lightweight transformer architectures specifically designed for edge deployment, analyzing recent advances in model compression, quantization, pruning, and knowledge distillation techniques. We systematically review prominent lightweight variants including MobileBERT, TinyBERT, DistilBERT, EfficientFormer, EdgeFormer, and MobileViT, providing detailed performance benchmarks on standard data

1 Introduction

This paper examines: Lightweight Transformer Architectures for Edge Devices in Real-Time Applications. Research question: How does the accuracy and inference latency of lightweight BERT models compare to distilled versions of larger language models like RoBERTa or DeBERTa when deployed for real-time Android malware detection on edge devices?.

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 5.0/10.

3 Results

14 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 5.0/10.

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.

References

- http://arxiv.org/abs/2601.03290v1
- http://arxiv.org/abs/2411.00907v3
- http://arxiv.org/abs/2502.15041v2