Scaling Inference Latency of SecLM Variants on Edge Devices vs. Cloud GPUs

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does the inference latency of SecLM variants scale with model size when processing multimodal inputs on edge devices compared to cloud GPUs. With the breakthroughs in deep learning, the recent years have witnessed a booming of artificial intelligence (AI) applications and services, spanning from personal assistant to recommendation systems to video/audio surveillance. More recently, with the proliferation of mobile. 3 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Edge Intelligence: Paving the Last Mile of Artificial Intelligence With Edge Computing. Research question: How does the inference latency of SecLM variants scale with model size when processing multimodal inputs on edge devices compared to cloud GPUs?.

2 Methodology

Systematic literature search across multiple databases yielded 12 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

12 papers retrieved. 3 claims extracted; 3 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
Billions of mobile and IoT devices are connected to the Internet, generating zillions bytes of data at the network edge.	✓	0.30
Edge computing pushes computing tasks and services from the network core to the network edge.	✓	0.27
Research on edge intelligence is still in its infancy stage.	✓	0.15

References

- https://doi.org/10.1561/2200000083
- https://doi.org/10.1109/jproc.2019.2918951
- https://doi.org/10.3390/fi16090329