How does the accuracy-throughput trade-off of Llama3-70B and Codestral-34B compare when deployed on heterogene

Assignee Research

May 29, 2026

Abstract

This paper proposes a neural architecture search (NAS) method for split computing. Split computing is an emerging machine-learning inference technique that addresses the privacy and latency challenges of deploying deep learning in IoT systems. In split computing, neural network models are separated and cooperatively processed using edge servers and IoT devices via networks. Thus, the architecture of the neural network model significantly impacts the communication payload size, model accuracy, and computational load. In this paper, we address the challenge of optimizing neural network architect

1 Introduction

This paper examines: Neural Architecture Search for Improving Latency-Accuracy Trade-off in Split Computing. Research question: How does the accuracy-throughput trade-off of Llama3-70B and Codestral-34B compare when deployed on heterogeneous edge devices (e.g., mobile, embedded) for HumanEval-hard, and what optimizations best maintain accuracy at minimal latency?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

16 papers retrieved. 2 claims extracted; 0 independently verified. Quality review score: 4.2/10.

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
BottleNet can improve end-to-end latency and reduce mobile energy consumption compared with cloud-based computation with	×	0.06
HW-NAS-Bench provides the computation latency and energy cost of all the network architectures in the search spaces of b	×	0.11

References

- http://arxiv.org/abs/2506.10461v1
- http://arxiv.org/abs/2302.10681v4
- http://arxiv.org/abs/2208.13968v1