Qeios

Peer Review

Review of: "GenAI at the Edge: Comprehensive Survey on Empowering Edge Devices"

Jianwei Hao[1]

1. Governors State University, United States

Summary:

This manuscript presents a survey of techniques and frameworks for enabling Generative AI (GenAI) deployment on edge devices. It is structured into three categories: Software optimization, Hardware optimization, and Frameworks. The authors also filter a broad set of recent works and propose a taxonomy that maps GenAI developments to resource-constrained edge computing environments.

Strengths:

1. This manuscript clearly organizes this topic into software optimization, hardware optimization, and frameworks, which is logical and helps readers navigate this vast topic. Each section includes detailed discussion.

2. The manuscript references some recent methods, many from 2023–2025, covering topics such as FlashAttention-3, EdgeQAT, QLoRA, etc.

Weaknesses:

1. The authors claim that "there is no dedicated survey on GenAI at the edge." However, the paper below has already explored this topic.

Guo et al., 2024, “A Survey: Collaborative Hardware and Software Design in the Era of Large Language Models”

doi.org/10.32388/47LPMG

2. This manuscript elaborates on some model compression methods. However, it lacks some topics like layer fusion, dynamic inference, etc., which makes it less comprehensive.

3. A table comparing methods to trade-offs is needed, e.g., accuracy loss and latency, limiting practical utility.

Declarations

Potential competing interests: No potential competing interests to declare.