Qeios

Peer Review

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

Hasanain Mohammed Manji Al-Rzoky[1]

1. Independent researcher

Public Review – Revised Version

The paper *"GenAI at the Edge: Comprehensive Survey on Empowering Edge Devices"* offers a timely and valuable overview of how generative AI (GenAI) models can be adapted to operate efficiently on edge computing systems. The topic is well-chosen and aligns with emerging research priorities in distributed intelligence and on-device AI.

Strengths:

The authors effectively categorize software optimization strategies into model compression, quantization, pruning, and distillation. For example, the discussion in Section 3 clearly outlines how *knowledge distillation* can be leveraged to reduce the computational load of large generative models without a significant loss in performance—a key requirement for real-time inference on edge devices. Similarly, the section on *runtime optimization frameworks* highlights how lightweight AI toolchains such as TensorRT and ONNX Runtime can be integrated to accelerate inference while maintaining low power consumption.

The paper also does a good job of summarizing **recent developments** such as the deployment of *transformer-based architectures* (e.g., GPT-style models) on constrained hardware, and the rise of *federated and collaborative edge training*, which are both highly relevant to the field's current direction. These elements make the survey timely and demonstrate the authors' familiarity with recent literature.

doi.org/10.32388/X7ZUUI

Declarations

Potential competing interests: No potential competing interests to declare.