IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, May 2025
Impact Factor: 7.67

Enabling On-Device AI for Android: A Dual-Mode System for Offline and Online Large Language Model Inference

Dr. Shailesh Bendale[1], Disha Raskar[2], Akshada Kadam[3], Shlok Jagtap[4], Shivraj Kore[5]
HoD, Department of Computer Engineering[1]
Students, Department of Computer Engineering[2-5]
NBN Sinhgad Technical Institute Campus, Pune, India

Abstract: *As large language models (LLMs) become increasingly important to upcoming applications and use cases, their dependence on cloud infrastructure raises concerns around latency, privacy, and accessibility. This paper presents a inference system that has two modes- offline and online modes for LLM execution. The offline mode leverages a quantized variant of DeepSeek-R1-Distill-Qwen-1.5B using Llama-RN on mobile hardware, while the online mode utilizes the cloud-based Gemma API. Our system, implemented on a consumer-grade smartphone, demonstrates the feasibility of on-device LLM inference without compromising accessibility or efficiency. We discuss implementation strategies, memory considerations, and trade-offs, contributing to the growing field of edge-native LLM deployment on mobile devices.*

Keywords: On-Device AI, Large Language Models, Hybrid Inference, Offline AI, Edge Computing, Llama-RN, Quantized Models, Mobile Inference, React Native, Gemma, Android