Vikas Mishra

vikasm3004@gmail.com +91 91401 12914 Pune, India LinkedIn GitHub

Profile Summary

Software Engineer in R&D working across Android development and AI, with experience building production-grade Android applications and end-to-end AI pipelines—from large-scale data processing and model evaluation to optimized on-device deployment on NPUs. Delivered solutions in collaboration with Qualcomm and NVIDIA, showcased at major technology events.

Skills

- **Languages:** Python, Kotlin, Java
- **AI / ML:** LLMs, GenAI, Retrieval-Augmented Generation (RAG), Agentic Frameworks, TensorFlow, PyTorch, Model Quantization (AIMET/INT8) i.e for On-device Inference (NPU)
- **Android Development:** Jetpack Compose, MVVM, StateFlow, LiveData, Room, WorkManager
- **Tools & Frameworks:** Git, Firebase, Android Studio, Olama, Postman

Work Experience

Associate Software Engineer Pune
Tech Mahindra – Makers Lab (R&D) March 2024 – Present

- Designed and deployed **AI-driven systems** including LLM pipelines and edge AI applications.
- **Indus – Educational LLM (Showcased at India AI Summit 2026):**
  - Built scalable **data pipelines** for large-scale corpus ingestion, cleaning, and preprocessing for LLM training.
  - Developed **evaluation and benchmarking framework** to compare model variants across accuracy and response quality.
  - Contributed to **LLM training workflow**, including dataset curation, model selection, and fine-tuning strategy.
  - Worked on **system-level design** for training and inference pipeline integration.
- **Fraud Call Detection – Edge AI (Showcased at IMC, in collaboration with Qualcomm):**
  - Designed and developed a **real-time fraud call detection pipeline** on Android, enabling on-device classification of live voice calls.
  - Built system to **capture voice call audio streams/packets** and process them in near real-time.
  - Built a **speech-to-text inference engine** to convert live call conversations into text using streaming transcription.
  - Integrated **on-device LLM-based classification** to detect fraudulent intent from transcribed conversations.
  - Optimized and deployed **quantized models (INT8) using AIMET** for execution on Qualcomm NPUs, ensuring low latency and efficient inference.
- **AutoVaani – Automotive AI Assistant (Showcased at Qualcomm Auto Day, in collaboration with Qualcomm):**
  - Developed a **voice-enabled Android automotive application** with real-time, offline AI capabilities.
  - Built an **end-to-end speech-to-intent pipeline**, converting user voice input into actionable commands using on-device models.
  - Integrated **edge AI inference** to enable low-latency responses in resource-constrained automotive environments.

- Designed system architecture for **offline-first interaction**, ensuring reliability without network connectivity.

- **Darpan Application (Showcased at India AI Summit 2026):**
  - Stabilized and optimized the application for large-scale demonstrations by resolving critical **memory leaks, ANRs, and concurrency issues**.
  - Improved runtime performance and responsiveness through **profiling, system-level debugging, and targeted optimizations**.
  - Designed and implemented the application architecture using **MVVM, StateFlow, and modular design**, enabling scalable state management and improved maintainability.

Projects

GenAI RAG Chatbot

- Designed and implemented an end-to-end **Retrieval-Augmented Generation (RAG) system** for context-aware question answering over domain-specific datasets.
- Built a **semantic retrieval pipeline** using dense embeddings and vector similarity search to fetch relevant context efficiently.
- Developed **top-K retrieval and ranking strategies** to improve response grounding and reduce hallucinations.
- Integrated LLM inference via **Groq API** to enable low-latency, real-time response generation.
- Applied **prompt engineering and response constraints** to improve factual consistency and control output quality.

Education

B.Tech in Information Technology 2019 – 2023
AKTU – Buddha Institute of Technology CGPA: 7.8

Certifications & Achievements

- Microsoft Certified: Azure Data Scientist Associate
- Best Team Award (2x) – Tech Mahindra
- Pat on the Back Award – Technical Excellence