Armin Mehran

Principal Machine Learning Engineer | Agentic AI, LangGraph & Edge Deployment
Toronto, ON, Canada

Email: a.mehran@gmail.com | Phone: +1-416-887-4647 | Linkedin: linkedin.com/in/arminmehran |
Portfolio: https://arminmehran.com | Github: https://github.com/Amehran

PROFESSIONAL SUMMARY

Principal Machine Learning Engineer (PhD) with 10+ years of experience architecting scalable AI systems. Specialized in **Agentic Workflows (LangGraph, CrewAI)** and **On-Device Inference (Gemini Nano)**. Proven track record of deploying secure, privacy-first ML pipelines for FinTech environments (TD/Banking standards). Expert in bridging serverless Python backends with edge clients to reduce latency and cloud costs.

TECHNICAL SKILLS

AI Backend & Agents: Python, FastAPI, CrewAI, LangGraph, LangChain, Pydantic, OpenAI API, Prompt Engineering.

On-Device AI & ML: Google Gemini Nano, MediaPipe, TensorFlow Lite, ML Kit, Quantization, CameraX, Computer Vision.

Core Engineering: Kotlin, Clean Architecture, CI/CD, Docker, System Design.

Cloud & DevOps: AWS, GCP, Docker, GitHub Actions, CI/CD, WebSockets/SSE.

PROFESSIONAL EXPERIENCE

Principal Mobile & AI Architect (Independent Consultancy) Toronto, ON | Dec 2023 – Present

Specializing in privacy-first AI architectures and bridging cloud agents with edge execution.

- **Scalable Multi-Agent Orchestration Engine (LangGraph):** Architected a stateful multi-agent system using **LangGraph** and FastAPI to handle complex reasoning tasks. Optimized memory management for long-context agentic loops.
- **Hybrid Cloud-Edge Inference Pipeline:** Engineered a privacy-preserving inference layer using **Gemini Nano** for local processing, falling back to cloud LLMs only when necessary. Reduced cloud token costs by 40%.
- **Key Initiative: Assistive Computer Vision Pipeline:** Built a real-time object detection system for visually impaired users using **TensorFlow Lite** and **CameraX**, optimizing for low battery consumption.
- **Research & Methodology:** Conducted performance benchmarking between Cloud LLMs (GPT-4) and On-Device SLMs to optimize trade-offs between accuracy, latency, and thermal constraints.
- **Enterprise Standards:** Maintained strict **CI/CD pipelines** and automated testing using **GitHub Actions**, mirroring enterprise-grade workflows to ensure code quality and maintainability.

Senior Mobile Application Developer Toronto, ON | Sep 2020 – Dec 2023

Tata Consultancy Services (TCS) | Client: Tier-1 Canadian Financial Institution

- Delivered critical features for a flagship banking application serving millions of users, adhering to strict OSFI security and compliance standards.
- **Legacy Modernization:** Re-architected legacy Java modules into modern **Kotlin** and **Jetpack Compose**, reducing code size by **55%** and significantly improving build times.
- **Leadership:** Led code reviews and enforced clean architecture principles (MVVM/MVI), resulting in **zero critical bugs** in production releases.
- Collaborated closely with iOS and Backend teams to ensure API contract alignment and feature parity across platforms.

Lead Mobile Application Developer Toronto, ON | Feb 2018 – Sep 2020

AP1

- Led a cross-functional team of 5 engineers to deliver the **apConnect SDK**, a scalable location service engine using Bluetooth Low Energy (BLE) beacons.
- Designed the Android SDK architecture using **Coroutines** and **Retrofit**, reducing partner integration time by **50%**.
- Spearheaded the migration of the "BeachLife" application to **React Native**, establishing a shared cross-platform codebase for easier long-term maintainability.

Software QA Developer Toronto, ON | May 2017 – Sep 2017

Cloud Constable

- Modernized "XFace," a legacy 3D avatar application, refactoring core logic using C++ and Visual Studio.
- Implemented automated testing pipelines for AI-based services, including a phishing email classifier, ensuring high precision/recall stability.

Mobile Application Development Consultant

Toronto, ON | Dec 2014 – May 2017

Armin Mehran

- Developed a voice-activated IoT controller using **Google DialogFlow (GCP)**, managing multi-turn dialogue states and intent recognition long before the current LLM wave.
- Built high-performance utility apps focusing on GPU rendering optimization and memory management.

CERTIFICATIONS & EDUCATION

- **The Complete Agentic AI Engineering Course** – Udemy (2025)
- **Meta Mobile Developer Professional Certificate** (2025)
- **Google Cybersecurity Professional Certificate** (2024)
- **AWS Certified Cloud Practitioner** (Expires 2026)
- **PhD. in Computer Engineering** – Azad University, Science & Research Branch
- **M.Sc. in Computer Engineering** – Azad University, Science & Research Branch
- **B.Sc. in Computer Engineering** – Azad University, Southern Tehran Branch