Aman Malhotra

+91-9464262941 | aman18.malhotra@gmail.com | LinkedIn | Github

Summary

Founding engineer who scaled consumer fintech to **200M+ monthly sessions** and **130M+ monthly payments** at CRED, then built production AI from zero at Naptick — **voice-to-voice agents**, **LLM orchestration**, **RAG**, and **embedded AI hardware**. Comfortable owning the entire stack: backend, cloud, mobile and firmware.

Core Expertise

AI Systems:	LLM orchestration, multi-agent systems, RAG, embeddings, vector search, conversational memory, tool-calling systems, conversational agents.
Infrastructure:	Distributed systems, cloud architectures, backend services, event-driven pipelines, release infra.
Full-Stack Systems:	Hardware + firmware + backend + mobile app integrations, end-to-end product ownership
Technologies:	Dart, Javascript, Node.js, NestJS, Python, Java, Swift, Docker, AWS, REST/gRPC APIs, protobuf, MongoDB, SQL, SQLite

Professional Experience

Founding Engineer, Naptick Aug 2025 – Present

- Shipped Naptick's first production **voice-to-voice AI agent** end-to-end across the mobile app and embedded hardware (**ESP32 / Raspberry Pi**).
- Optimized the streaming pipeline to achieve **sub-second voice-to-voice latency**.
- Built **audio encoding and compression pipelines** engineered to run on low-RAM embedded devices, enabling on-device sound processing within tight memory budgets.
- Optimized system components to run within a **∼300KB embedded RAM footprint**, using careful memory management, streaming, and buffer-reuse strategies to keep the device stable under severe resource constraints.
- Built the **LLM orchestration layer** with multi-agent routing, **RAG over sleep data using embeddings and vector search**, and conversational memory - enabling the agent to retrieve relevant health context.
- Designed and implemented **tool-calling pipelines** that let AI agents trigger product capabilities and background workflows.
- Owned the full stack — **firmware, backend, cloud, and mobile** — taking AI features from hardware signal through to user-facing insight.
- Implemented observability for AI workflows using **CloudWatch logging, event-driven pipelines, and Mixpanel analytics** to monitor model behavior and production reliability.
- Took major application redesigns and backend rewrites from concept to production **within days**, keeping pace with rapidly shifting product direction.
- Built and maintained **release pipelines and CI/CD workflows** across firmware, backend services, and application deployments.

Senior Software Engineer, CRED Mar 2021 – Aug 2025

- Built and scaled backend systems powering **200M+ monthly active user sessions** across core consumer experiences.
- Delivered infrastructure supporting **130M+ monthly payment transactions** with strong reliability requirements.
- Led migration of CRED's mobile codebase from **three repositories to a unified monorepo** used by **∼80 engineers**, improving development and QA velocity.
- Built custom migration pipelines to consolidate repositories totaling **∼500GB** while preserving full git history and LFS assets.
- Introduced **Protocol Buffers at the app interface layer**, which later expanded across multiple business lines to improve API efficiency and contract stability.
- Drove hiring and internal knowledge sharing by conducting **∼80 technical interviews** and leading sessions on **Flutter multi-engine architecture, Protocol Buffers, and secure handshake systems**.
- Took ownership of delivering multiple product initiatives end-to-end, working closely with product and design teams to keep launches on timeline.
- Initiated early integration work with **Shorebird** to explore modern mobile deployment and runtime patching capabilities.

Education

Bachelor of Technology (B.Tech), Computer Science and Engineering 2017 – 2021
Punjabi University, Patiala