Amir Rezaei

AI / LLMOps Engineer · Embedded AI & Firmware

+98 992 705 3144 | linkedin.com/in/amir-rezaei | github.com/OD-V | odi-v.com | Tehran, Iran

Summary

AI / LLMOps engineer specialised in fusing firmware with AI. Builds systems where microcontrollers, sensors, displays, microphones and cameras feed **production LLM agents** — and where the agents reach back to drive the device in real time. Hands-on across the whole stack: **ESP32 firmware in C++**, device protocols, event-sourced backends, retrieval memory, realtime voice, multi-model routing, evaluation and cost governance. Previously a full-stack developer on a **multi-chain crypto exchange**.

Skills

- **Embedded & Firmware:** C/C++, ESP32 / ESP32-S3, FreeRTOS dual-core tasking, PlatformIO, Arduino-ESP32 / ESP-IDF, LovyanGFX, PSRAM / IRAM / NVS, I^2C · SPI · I^2S, BLE GATT, TLS on MCU, Wi-Fi provisioning, WebAssembly ports
- **AI & LLMOps:** multi-provider routing (OpenRouter, Anthropic, OpenAI, Gemini), prompt caching, fallback chains, cost tracking, tool-calling agents, MCP servers, blind A/B evaluation, LLM-as-judge
- **RAG, Voice & Vision:** PostgreSQL + pgvector, embeddings, retrieval ranking, Gemini Live speech-to-speech, Deepgram, ElevenLabs, VLM video analysis
- **Backend & Web:** Node.js, TypeScript, Python, PHP / Laravel, React, Next.js, Redis, WebSocket, nginx, systemd, Cloudflare, Linux, Git
- **Web3:** web3.js, EVM (Ethereum, Base, Optimism), Tron, Solana, wallet services, deposit / withdrawal flows

Work Experience

• ODI-V *Jul 2026 – Present*

Founder — Lead AI, Firmware & Hardware Engineer Tehran, Iran · On-site

- Building **Melo, the first On-Desk Intelligence device**: an AI desk companion with long-term memory that follows up on what matters to its user. Lead design, hardware, firmware and the AI layer end to end.
- Architected **dual-core ESP32-S3 firmware** (124 C++ modules, 600+ commits): rendering pinned to one core at a steady **~38 fps**, networking and TLS telemetry on the other — **zero panics** in post-merge soak runs.
- Built a **procedural face engine** with tweening, lip-sync and **95+ animated behaviours** (talk moods, sleep, wake, games); compiled the same C++ engine to **WebAssembly** so device and web render identical faces.
- Engineered under hard memory limits: **2.5–3 KB TLS heap margin**, instruction-cache layout tuning worth ±5 fps, PSRAM sprite stores and IRAM hot loops; contributed an **upstream fix to LovyanGFX** (PR #939).
- Closed the **device ↔ AI loop**: soil, temperature, humidity, light and capacitive touch sensors → on-device brain with hysteresis → telemetry → LLM agent → per-user parameters pushed back into device NVS; **realtime WebSocket link** syncs the face with chat and voice.
- Shipped secure provisioning: **captive-portal and QR Wi-Fi onboarding**, pinned root-CA TLS, pairing codes with hashed one-time device secrets.
- Designed the AI layer: **event-sourced agent core**, **pgvector long-term memory** (77.2% vs 49.9% F1 tagging), **Gemini Live voice calls**, a proactive follow-up engine, and blind model evaluation with prompt-cache cost control.
- Led hardware generations: ESP32-S3 board selection, **pin maps and power budgets**, I^2S microphone and speaker, OV2640 camera path, LED status ring, **3D-printed PETG enclosures** with snap-fit joints and heat stakes.
- Operate **25+ production services** and **3,200+ commits** on the platform, leading a team of human and AI coding agents.

• **PartoMali** *Mar 2025 – Jul 2026*

Full-stack Developer — Pump-Ex Crypto Exchange Tehran, Iran

- Developed the **frontend and backend** of the Pump-Ex exchange: **React / Next.js** client, **Laravel (PHP)** API and admin panel, and a **Node.js wallet service** built on **web3.js**.
- Delivered **multi-chain support across Layer-1 and Layer-2 networks** — Ethereum, Base, Optimism, Tron and Solana — covering wallet generation, balances, deposits and withdrawals.
- Worked on a **chain-registry design** with one provider interface per chain family, per-network gas settings for L2s, and a **Redis-cached market price service**.

Projects

• **be-zang — Autonomous AI Phone Agent** *Telephony · Realtime*

- Places **real phone calls** toward a goal, steered **mid-call** by a supervising reasoning model; audio bridged from a physical handset.
- STT benchmark on **77 labelled Persian calls** (86% vs 72%) decided the production speech engine.

• **Smart-Ring Health Platform** *BLE · Wearable · IoT*

- **Reverse-engineered the BLE GATT protocol** of a retail smart ring (heart rate, SpO_2, temperature, steps) and replaced its vendor app with our own ingestion and API.
- Rule-based **alert engine** for family and organisation watchers, covered by end-to-end tests.

• **Edge Vision — Shoplifting Detection** *VLM · Edge*

- Blind-tested **30+ vision-language models** on CCTV clips; found CPU-only options at **$0.0002 per clip** and designed a **three-tier path**: on-device reflex → light server model → frontier model.

• **Melo Vision — Camera Pipeline for an AI Device** *Firmware · Vision · Realtime*

- Designed a **binary frame protocol** (12-byte header) from an ESP32-S3 camera to the server, with presence, gaze-follow and face-mood replies driving the device face.
- Three-layer split: **reflex on the device**, light model on the server, frontier vision model only on demand; a browser simulator throttles frame rate to mimic weak boards.

• **Production AI Pipelines** *Automation · Content · SEO*

- **Live lecture notes** from streaming speech, an **animated reel engine** with multi-voice TTS and frame QC, and an **SEO automation service** with safety guards.

• **Hardware Concepts** *Electronics · Mechanical*

- Sensor-only **elderly wristband** feeding an AI companion, and a portable **solar-panel cleaning robot**, both designed from locally available modules.

Education

• **Islamic Azad University** — *B.Sc. in Computer Engineering* *In progress*