

Shreyash Dubey

📍 Bangalore, India, open to relocating to the EU | ✉ shreyashlrn@gmail.com
🌐 shreyashdubey.com | 🔗 linkedin.com/in/dubeyshreyash | 🐙 github.com/shreyashdubey

AI engineer shipping production agents into regulated, customer-facing workflows: tool-calling agents, retrieval and memory, eval harnesses, tracing and cost control.

Four years owning systems end to end, often as the sole engineer, across healthcare, wearables and e-commerce.

SKILLS

GenAI & Agents: LLM tool calling and API integrations, prompt engineering, retrieval-augmented generation, conversational memory, vector search (Qdrant), deterministic analytics under an LLM narrator

AI Quality & Monitoring: Eval harnesses on frozen data with statistical regression tests, guardrails, request tracing, per-call cost accounting, prompt caching

Languages: Python, TypeScript, SQL, C, C++, Swift

Cloud & Backend: AWS (EC2, RDS, SSM), FastAPI, Django, Node.js, REST, SSE, PostgreSQL, SQLAlchemy/Alembic, MongoDB, Linux

Devices: Bluetooth LE sync, NFC sensor reads, biometric signal pipelines (PPG, HRV, sleep), drone firmware

EXPERIENCE

AI Engineer

April 2026 – Present

Muse Wearables — smart ring hardware

Bangalore, India

- Sole engineer on **Rio**, the AI companion for the Ring One smart ring, in production since July 2026: one **tool-calling agent over 40 typed tools**; the model never writes SQL. Python/FastAPI on AWS EC2, PostgreSQL on RDS.
- Deterministic detectors** (personal baselines, anomalies, streaks, correlations with minimum sample sizes) find the patterns and the LLM only narrates them; every figure in a reply is checked against the tool data.
- Eval harness:** 232 cases run 3x against a frozen database and pinned clock, assertions gate, McNemar's test flags regressions; ~1,300 tests. Multi-provider failover and a hard per-message cost cap.

AI Engineer

Dec 2025 – April 2026

August AI — for Manipal Hospitals, on site

Bangalore, India

- Sole owner of a **production patient-facing agent** at **5,000 messages and 100 completed bookings per day** for a major Indian hospital chain, in a regulated healthcare setting.
- Embedded with clinical and operations staff: turned their booking workflows into agent specifications, then into **live tool calls** against hospital systems (booking, cancellation, patient-ID lookup) with constrained slot search over doctor, branch, specialty and time.
- Absorbed a **6–7x load swing** (250–300 daily signups to 2,000 during marketing campaigns) with no rewrite: profiling, caching and concurrency work on the hot path.

Full Stack and AI/ML Engineer

June 2024 – Dec 2025

2070 Health (W Health Ventures studio) — four portfolio companies

Remote

- Built the shared **EHR backend** running both ElevateNow and Nivaan Care, integrating Orange Health labs, Tata 1mg pharmacy and WhatsApp booking — one platform, multiple internal customers with conflicting requirements.
- AI call-quality service** at Everhope Oncology: transcription pipeline scoring calls against a clinical rubric. BabyMD landing pages from 4 days to under 4 hours with a page generator.

Backend Developer (Freelance), Folojet (*Nov 2023 – April 2024, San Francisco client*) — shipped a public Shopify app solo in Node.js: order and fulfilment sync, FedEx shipping, merchant tooling.

Software Developer

Sept 2022 – Oct 2023

Muse Wearables — hybrid smartwatch and glucose sensor

Bangalore, India

- iOS app: the **Bluetooth sync layer** (packet framing, retry and reconciliation over an unreliable BLE link), **NFC glucose-sensor reads**, sleep and activity scores.
- Built a **driver drowsiness detection** prototype for Renault.

Drone Firmware Developer (Intern), Inoneticx (*2018 – 2021, Lucknow, India*) — LiDAR obstacle avoidance and vision-based object tracking for autonomous drones, tuned to on-board compute and memory limits.

SELECTED WORK & EDUCATION

- Samvaad Cafe** — offline-first point of sale: event outbox reconciled with PostgreSQL, thermal printing from the browser, integer-paise money; 200 tests. **HLD Gym** — live 51-chapter system design book and redraw-from-memory practice gym.
- Atlas** — LLM coach with long-term memory (per-user Qdrant recall over PostgreSQL, request tracing, per-call cost); built the backend and memory pipeline with a collaborator on the clients. **B.Tech Computer Science**, BBDNIIT Lucknow, 2022.