

NILESH S. PATIL

Belgaum, Karnataka, India · +91 84314 96045 · nilesh@nileshpatil6.com

LinkedIn: www.linkedin.com/in/nileshpatil6 · GitHub: github.com/nileshpatil6 · Portfolio: nileshpatil6.com

PROFILE

Final-year AI & Data Science engineering student with **18 merged pull requests** in the production codebases of **Google, Microsoft, Apple, Uber, OpenAI, Meta, and Hugging Face**. 10x Hackathon winner, Grand Finalist at the **Meta × Scaler PyTorch Hackathon** and **Top 30 nationally** in ISRO's IRO-C-U 2026 autonomous robotics challenge. Co-founded an **Rs. 2L government-funded** healthcare startup and delivered freelance production software for clients across India and Japan. Focus: efficient ML systems, agentic AI, autonomous robotics, and full-stack systems.

EXPERIENCE

Open Source Contributor

2025 – Present

18 merged pull requests across Google, Microsoft, Apple, Uber, OpenAI, Meta, Hugging Face, SciPy

Remote

- **Apple** – ml-explore/mlx #4173: kept double precision for Python floats in float64 operations in Apple's ML array framework.
- **Uber** – uber/h3 #1224: fixed an out-of-bounds read in the H3 geospatial index core (C), reproduced under AddressSanitizer and closed with a bounds check plus regression test.
- **Google** – grpc/grpc-go #9345: fixed ALTS credentials `Clone()` silently dropping the bound access token; further merged fixes in google/osv-scanner (+**711 LOC**), google/pprof, and google/go-containerregistry.
- **Microsoft** – microsoft/garnet #2070: fixed a disposal race in the Redis-compatible cache-store's object allocator; two further PRs merged in microsoft/pyright.
- **OpenAI / Meta / Hugging Face / SciPy** – fixed silent error suppression in the OpenAI Agents SDK (#2931), Hydra sweep-abort callbacks, a transformers-serve endpoint, and a SciPy optimizer hang (#25912).

Co-Founder and Engineering Lead, MediAssist AI

2024 – Present

Government of Karnataka NAIN 2.0 Grantee, Rs. 2L grant

Belgaum, India

- Architected 2 mobile apps (patient and doctor) on 1 unified Supabase backend with row-level security, Edge Functions, and live subscriptions, serving 3 user roles.
- Built a voice-to-voice multimodal AI loop using Whisper, GPT-4o, and OpenAI TTS, plus prescription OCR with GPT-4o Vision, medication reminders, and Emergency SOS.

Freelance Full-Stack and AI Engineer

2023 – Present

Independent Clients across India, USA, and Japan

Remote

- Delivered 7 live production apps across real estate diagnostics, interior design, car rentals, enterprise facility management, and a Japanese construction photo-register SaaS (Flutter + Next.js + Supabase monorepo).
- Shipped client products currently live: sumaimatch.jp, prasanhom.in, unyfiny.com, cmnservicesindia.com, ifuturetech; plus own products keepiai.com and productcompareai.com.

PROJECTS

PromptInject Environment – Adversarial RL for LLM Security – PyTorch, GRPO, LoRA, Gemma 3

Meta × Scaler Finalist

- OpenEnv environment where a 5-agent GRPO attacker swarm trains against an online-learning LoRA defender; over 30 A100 episodes defender accuracy rose **47.1% to 92.8%** and attacker evasion fell **52.9% to 7.2%**. Also authored MindRead, a Theory-of-Mind RL environment; both live on Hugging Face Spaces.

Autonomous GNSS-Denied Micro-UAV – ISRO Robotics Challenge – Python, OpenCV, YOLOv8, ArduPilot

Top 30 in India

- Top 30 nationally in URSC/ISRO's ASCEND challenge; wrote a custom visual-inertial SLAM stack for flight without GNSS, ArUco-marker precision landing, and a RealSense depth + YOLOv8/SAM2 tracking pipeline on a Raspberry Pi companion computer with a ground-station telemetry server, validated across logged autonomous missions.

Keepi – AI Second Brain – Flutter, Isar, Gemini Live, Supabase

Google Play

- **300+ users within 20 days of launch**. Offline-complete Android app (11 modules) with **on-device semantic search** over notes, ranking Gemini embeddings by cosine similarity in a background isolate to keep the UI jank-free, and a realtime voice assistant over Gemini's **bidirectional Live WebSocket** that drives device actions via function-calling; patched 2 upstream Dart packages and proxied all model traffic so no API key ships in the APK.

AgentRAG – Long-Context Retrieval with a Council of Agents – Python, Gemini 2.5 Flash, ChromaDB, asyncio

GitHub

- Standard RAG concatenates every retrieved chunk into one oversized prompt, where attention cost grows superlinearly and the model averages its sources rather than reasoning over each. AgentRAG instead gives **every chunk its own agent with an isolated context window**: agents run in parallel, an orchestrator runs a cross-examination round where they challenge each other's findings, and a synthesis agent emits the final cited answer. Total evidence reasoned over **exceeds any single context window** while each call stays short.

TECHNICAL SKILLS

Languages: Python, C, C++, TypeScript, JavaScript, HTML/CSS, Dart

Frameworks and Data: Flask, Prisma, Supabase

Cloud and Tools: AWS, Google Cloud, Vercel, Claude Code

HONORS & ACTIVITIES

- **Hackathon Prizes:** GDG, CodeBharat (Rs. 50,000), ONEST (Rs. 25,000), 5 others; NASA SpaceApps 1st Place (Local).
- **Meta × Scaler PyTorch Hackathon:** Grand Finalist (2026). **IIT Bombay Campus Ambassador** (2024).

EDUCATION

S.G. Balekundari Institute of Technology

2023 – 2027

B.E. in Artificial Intelligence & Data Science (Final Year)

Belgaum, Karnataka, India