HARDEEP SINGH

AI Engineer | New Delhi, India | | hardeep.cv | hardeepsindia@gmail.com | +91 97175 17181 | github.com/rav4nn

PROFILE

Self-taught AI engineer with **2,000+ organic users** across deployed products and **60+ GitHub stars** on open-source tools. Builds and ships end-to-end LLM-powered applications using Python, FastAPI, Next.js, and the full AI stack — RAG pipelines, agentic systems, vector search, prompt engineering, and multi-provider LLM integration (Anthropic, OpenAI, Gemini, DeepSeek). Co-built a crisis-response platform serving **15 Indian states in 48 hours**.

EXPERIENCE

Stealth SaaS — Sole Engineer (Freelance) *2026 – Present*

End-to-end ownership of a full-stack SaaS product for a UK-based client — architecture, build, and production deployment. Scope includes AI-powered features and agentic workflows.

Self-Directed — AI Engineer (Independent) *2025 – Present*

Built and shipped 0 to 1 AI products with organic growth and real user traction. Products include coffeecoach.app (**65 daily active users**), Splitwala (**2,000+ users across 117 WhatsApp groups**), youtube-rag-scraper (**60+ GitHub stars**), FluxRAG, and buildinpublic-x. Spent 2023 to 2024 learning the stack before shipping the first product in early 2026.

CovidWin — Operations Lead *Apr 2021 – May 2021*

- Co-built a COVID-19 resource platform from zero to full operations in **48 hours** during India's second wave — covering **15 states, 50 cities**, and 10,000 verified life-saving resources.
- Led volunteer coordination across 8 states, managing **4,000 volunteers** from partner organisations.
- Built a 15-minute automated data sync pipeline using Google Sheets API with deduplication and multi-source aggregation.

Digital Marketing — Freelance *2018 – 2024*

Independent client work across digital marketing and process engineering. Transitioned out to pursue software and AI engineering full time.

PROJECTS

Coffee Coach coffeecoach.app

Python · FastAPI · Next.js · React · PostgreSQL · Docker · LangChain · LLM APIs

- Full-stack AI coaching app for specialty coffee brewers — personalised feedback from an AI coach based on brew history and taste data, using a RAG pipeline with agentic feedback loops.
- **65 daily active users** acquired organically via Twitter/X and Reddit with zero paid promotion.
- End-to-end production deployment: Next.js on Vercel, FastAPI on Hetzner VPS with Docker, Nginx, and Certbot.

FluxRAG github.com/rav4nn/flux-rag

Python · RAG · Hybrid Search · Re-ranking · FastAPI · Embeddings · Vector Search

- RAG evaluation harness that inverts the typical workflow — benchmarking first, optimisation second. Supports **10 file formats** with parameter sweeps across chunking strategies and **8 embedding models**.
- Generates ranked benchmark reports tracking **latency, cost per query, and hallucination rate** across every configuration — so retrieval quality and cost tradeoffs are measured, not assumed.
- Async FastAPI server with hybrid search and re-ranking. Built to validate production RAG setups before shipping.

YouTube RAG Scraper github.com/rav4nn/youtube-rag-scraper

Python · LangChain · FAISS · RAG · Embeddings

- Pipeline that bulk-scrapes YouTube channel transcripts, processes and chunks them, embeds into a FAISS vector store, and exposes a semantic search interface over the resulting knowledge base.
- **60+ GitHub stars and 13 forks** gained organically — most-starred project in the portfolio.
- Full RAG stack: ingestion, chunking, embedding, retrieval, and generation. Handles retrieval quality out of the box.

buildinpublic-x github.com/rav4nn/buildinpublic-x

Python · GitHub Actions · LLM APIs · Multi-provider (Anthropic, OpenAI, Gemini, Groq, DeepSeek)

- GitHub Action that reads commit history and README, generates a build-in-public thread via LLM, and auto-posts to X and Bluesky on a cron schedule. No backend, no database — runs entirely inside GitHub Actions.
- LLM provider swappable via a single environment variable. Costs **~$0.01 per post**.

Splitwala github.com/rav4nn/splitwala-webjs

Node.js · whatsapp-web.js

- WhatsApp chatbot for group expense splitting — /split, /paid, /balances commands. No app to download, no account to create. Lives inside the existing WhatsApp group.
- Live across **117 groups with 2,000+ active users**. Built on the principle that the best UX is the one that gets out of the way.

SKILLS

Languages: Python, JavaScript (Node.js), HTML/CSS

AI / ML: LangChain, LangGraph, FAISS, RAG Pipelines, Agentic Systems, Prompt Engineering, Embeddings, Vector Search, Hybrid Retrieval, Re-ranking, LLM Evaluation

LLM APIs: Anthropic (Claude), OpenAI, DeepSeek, Gemini, Groq

Backend: FastAPI, REST APIs, PostgreSQL, Next.js, React, TailwindCSS, Docker, Nginx, MCP Servers

Deployment: Hetzner VPS, Vercel, Cloudflare Pages, Railway, Fly.io, Git/GitHub, GitHub Actions

EDUCATION

Indian Institute of Technology (IIT) Delhi — Diploma, Chemical Engineering *2018*