HARDEEP SINGH

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

Profile

AI engineer with two concurrent contracts and **2,000+ organic users** across deployed products and **60+ GitHub stars** on open-source tools. Currently contracting at Squidgy AI (4142 Ltd, UK) and co-founding Coffee Coach, an AI coaching app grown to 65 daily active users with zero paid spend. Builds 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

4142 Ltd / Squidgy AI — AI Software Engineer *Apr 2026 – Present*

Full-time contract, remote, UK hours | squidgy.ai

- Contracted as an AI software engineer at Squidgy AI, a multi-agent AI product by 4142 Ltd (UK).
- Working within the engineering team on AI-powered features and agentic systems, full-time remote mapped to UK working hours.

Stealth SaaS — Sole Engineer *Apr 2026 – Present*

Freelance contract, UK-based client, NDA

- 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. Client and product details under NDA.

Coffee Coach — Founding Engineer *Jan 2025 – Present*

Self-employed | coffeecoach.app | Delhi, India

- Built Coffee Coach 0 to 1: a full-stack AI coaching app for specialty coffee brewers with personalised feedback using RAG pipelines and agentic feedback loops.
- 65 daily active users acquired organically via Twitter/X and Reddit with zero paid promotion.
- Now co-founding and productizing with a product partner, building towards commercial launch.
- Stack: Next.js on Vercel, FastAPI on Hetzner VPS with Docker, Nginx, Certbot, PostgreSQL.

CovidWin — Operations Lead *Apr 2021 – May 2021*

Full-time, remote

- 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.

Freelance Digital Marketer — Process Automation *Apr 2018 – Dec 2024*

Freelance, Delhi, India

- Independent client work across digital marketing: SEO, content, and campaign management. Later years included lightweight process automation, leading to a full pivot into software and AI engineering in 2025.

Projects

FluxRAG *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.
- Async FastAPI server with hybrid search and re-ranking. Built to validate production RAG setups before shipping.

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.

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 *Node.js · Telegram Bot API*

- Telegram bot for group expense splitting: /split, /paid, /balances commands. No app to download, no account to create.
- 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*