RAHUL AGARWAL

AI Engineer & Full-Stack Developer

Bengaluru, India • +91-9461203345 • rahulagaral1@gmail.com • rahul-agarwal.in
linkedin.com/in/rahul-agarwal-ai • github.com/rahulagarwal18 • dev.to/rahul_agarwal18

PROFESSIONAL SUMMARY

AI Engineer & Full-Stack Developer who ships production AI software end to end. Currently the sole engineer at **Aryan News Agency**, building its AI-powered web platform and in-house **accounting system with an LLM assistant** from requirements to deployment. Builds open-source AI developer tools, including two published npm packages with **1,000+ downloads in their first two weeks**, an **MCP server** for AI agents, and offline local-LLM tooling. Strong in Python, FastAPI, Node.js, Next.js, TypeScript, RAG pipelines and cloud deployment. UN Millennium Fellow and National Science Day 1st Place winner.

TECHNICAL SKILLS

Languages:	Python, TypeScript, JavaScript, SQL, C++, Java
Web & Frontend:	Next.js (App Router, SSR), React 19, Vite, Tailwind CSS, HTML5, CSS3, REST APIs
Backend & Cloud:	FastAPI, Flask, Node.js, PostgreSQL, SQLite, AWS (EC2, S3), Vercel, Docker, CI/CD, GitHub Actions
AI & LLM Systems:	Generative AI, LLMs, AI Agents, MCP, RAG, LangChain, Prompt Engineering, Embeddings, Vector Search (FAISS)
LLM Platforms:	Claude API (Prompt Caching), OpenAI API, Llama 3 via Groq, llama.cpp (local GGUF inference), OpenCV
Tools & Practices:	Git, npm package publishing, Playwright, Slack Bolt, Zoho CRM APIs, Technical SEO, Core Web Vitals, Linux

PROFESSIONAL EXPERIENCE

Aryan News Agency — AI Engineer & Full-Stack Developer **Jul 2026 – Present**

- Own the agency's **AI and software engineering end to end as the sole engineer**, from AI roadmap and architecture to deployment and maintenance of all internal and public-facing systems.
- Architected a **dynamic, database-driven website** with an **LLM-powered content pipeline** that auto-generates summaries, tags and SEO metadata, so staff publish updates in real time with no developer involvement.
- Built a **custom accounting application** from scratch that digitized core financial operations (customer accounts, billing & invoicing, payments, outstanding dues, reports), with role-based access and validation to keep financial data accurate and secure.
- Integrated an **AI assistant into the accounting app** that answers plain-English questions about the books (*"which accounts have dues over 30 days?"*) using LLM tool calling over guarded, read-only queries, and drafts payment reminders.

InferWorks — Founder's Intern & Full-Stack Developer **Nov 2025 – Feb 2026 | Bengaluru (On-Site)**

- Built **LLM chatbots and agent workflows** with LangChain and Llama 3, cutting operational query turnaround by **~60%**.
- Built REST API integrations across **Zoho CRM** and backend datastores, automating lead capture and sync for **5,000+ records**.
- Deployed production web apps on **AWS** (hosting, DNS, SSL) with **CI/CD via GitHub Actions** for zero-downtime releases. Also ran technical SEO (JSON-LD schema, Open Graph, metadata) across company web properties.

OPEN-SOURCE & AI PROJECTS

Unmuzzled — Private, Offline AI Chat [GitHub] *TypeScript, Node.js, llama.cpp, GGUF*

- Published an npm CLI (**npx unmuzzled**) that runs open-weight LLMs **100% locally**, no API key or cloud. **500+ downloads in week one**.
- Shipped GPU-accelerated inference via node-llama-cpp (CUDA / Vulkan / Metal with CPU fallback), streaming Markdown responses, a live reasoning view, edit-and-regenerate, and searchable conversation history.

ClaudeCut — Claude API Cost Optimizer [GitHub] *Python, FastAPI, Anthropic API, npm*

- Built a drop-in local reverse proxy that places prompt-cache breakpoints correctly and auto-switches models by task, **cutting Claude API & Claude Code costs by 40–70%** with no code changes. Ships with a live cost dashboard; **500+ downloads in two weeks**.

LinkedIn Guardrail MCP Server [GitHub] *TypeScript, MCP (JSON-RPC), Playwright*

- Built an account-safe **Model Context Protocol** server that lets AI agents (Claude Desktop, Cursor, Windsurf) search jobs, inspect profiles, post and message on LinkedIn. It is safe by design: no stored cookies, human-paced actions and daily safety quotas.

NeuralDocs — Real-Time Document RAG System [GitHub] *Python, Flask, LangChain, FAISS, Llama 3*

- RAG app for cited Q&A across large PDFs, pairing FAISS with Groq-hosted Llama 3 for **sub-second** responses.

Also built: TicketSlayer AI (autonomous Slack agent that turns support chats into GitHub Issues via Llama 3) • AI Resume Screener (Next.js SaaS with vector-search candidate ranking) • OMR Vision Pipeline (OpenCV, 99% grading accuracy, 100+ sheets in <5s) • GitPulse (GitHub analytics storyteller) • DeepDecompile (AST decompiler for minified web bundles) • Interactive terminal resume (**npx rahulagarwal18**)

HONORS, RECOGNITION & WRITING

- **UN Millennium Fellow (2025–2026):** Selected for a global leadership program convened by United Nations Academic Impact (UNAI).
- **National Science Day Winner:** 1st Place (2025) for Smart Campus Navigation System, and 2nd Place (2024) for Biometric Verification.
- **Technical Writer:** Published **25+ articles** on dev.to covering MCP, LLM cost optimization, RAG architectures, local AI and React 19.

EDUCATION

CHRIST (Deemed to be University) — Bachelor of Computer Applications (BCA) **Bengaluru | 2023 – 2026**