Sahil Saini

Full-Stack Software Engineer | AI, LLM & RAG Systems | React, Node.js, Python, AWS
thisissahilsaini@gmail.com | linkedin.com/in/realsahilsaini | sahilsaini.tech
Indian National | **Former UAE Resident (Emirates ID)** | **Ready to Relocate to the UAE** | **Visa Sponsorship Eligible**

Professional Summary

Full-Stack Software Engineer building production **LLM and RAG systems** end to end with React, Next.js, TypeScript, Node.js, Python, PostgreSQL and AWS. Shipped a **production-grade AI RAG chatbot** and a **multi-tenant fintech SaaS** solo; now leading the web platform of a **UAE-hosted, Arabic-ready AI document SaaS** for a UAE client. Experienced with clients across the **UAE, USA, and global markets**; seeking a UAE-based role.

Experience

Full-Stack Engineer | *Unico Connect, Hybrid, Mumbai* **Feb 2026 – Present**

- Leading the web platform (customer PWA, Admin Portal, marketing site) of a UAE client's **AI document-intelligence SaaS**, where AI reads business documents, tracks their deadlines and answers **RAG search** questions; hosted on **AWS UAE**, Arabic/RTL-ready, launching Dec 2026 with a 4-person team.
- Delivered a production multi-tenant **equity-management fintech SaaS** (7 instrument types, native e-signatures, automated vesting, real-time fully diluted cap table across 3 portals) integrated with **CorpTools, Stripe, SendGrid, GCS**; reduced Delaware C-Corp incorporation turnaround from weeks to under **48 hours**.
- Made the fintech SaaS feel **instant across all 3 portals**: audited **221 save actions** and rebuilt the data layer on **React Query** (optimistic updates with rollback, server-confirmed cache writes, prefetching), removing stale-data flashes and spinners.
- Used advanced **Claude Code** workflows (agentic sub-agents, custom slash commands, MCP servers, hooks, `CLAUDE.md` context engineering) to replace a projected 3–4 developer team on the fintech SaaS while maintaining production-grade quality.

Software Developer | *Ahy Consulting, Mumbai* **Feb 2025 – Feb 2026**

- Engineered a **production-grade AI RAG chatbot** using open-source LLMs and custom pipelines with zero cloud dependency.
- Achieved a **98% Google Lighthouse performance score** through architectural and performance optimizations.
- Built scalable interfaces and services while **collaborating with international clients** across the **UAE**, USA, and global markets.

Full Stack Developer Intern | *DevTown, Remote, Mumbai* **Dec 2021 – Mar 2022**

- Developed and optimized backend systems for improved performance and scalability; performed comprehensive API testing with Postman; designed responsive, modern user interfaces with a performance-first approach.

Projects

AI Document Intelligence SaaS (UAE Client) | *React 19, TypeScript, Next.js 16, Better Auth, Claude on Bedrock, pgvector, AWS*

- Integrating an async **LLM document pipeline** (Claude on **Amazon Bedrock**, SQS workers) and **pgvector RAG search** into the product: typed and voice queries (speech-to-text), AI answers with citations **streamed over SSE**, and live per-file processing status with retry and password unlock.
- Building 3 apps (installable **PWA**, staff Admin Portal, Next.js marketing site) in a Turborepo monorepo on **AWS UAE (me-central-1)**, **Arabic/RTL-ready** from day one (i18next, build-checked logical CSS); a strict CSP lets the product apps connect only to the UAE-hosted API and S3.
- Co-designing the **REST API contract** with the backend team (typed OpenAPI client) and building the auth, access and upload flows on it: **Better Auth** HttpOnly cookie sessions, server-resolved **RBAC**, and **presigned direct-to-S3 uploads** of up to 200 files with in-browser image compression.

Multi-Tenant Equity Management SaaS | *Next.js 16, NestJS 11, PostgreSQL, Prisma, Redis, GCP, Claude Code*

- Engineered secure backend with **schema-validated APIs**, typed DTOs, structured audit logging, DB-backed **RBAC (13 granular permissions)** with Redis-cached lookups, JWT refresh rotation with blacklist revocation, TOTP 2FA, and AES-256 field encryption.
- Designed a **cache-first architecture**: a central query-key factory, shared optimistic and refresh helpers, and **10 endpoints** reshaped to return the same data as their GETs, so every save's response reaches each affected screen before navigation.
- Built an **LLM-powered document templating engine** that auto-injects entity variables into legal agreements via a post-commit side-effect pipeline; shipped solo using an AI-first workflow, compressing a 22-week roadmap (3–4 dev team) by an estimated **70%**.

DermAI | *Python, Flask, Streamlit, AstraDB, OpenAI Embeddings*

- Built an AI dermatologist chatbot achieving **97% diagnostic accuracy** using LangChain, Chainlit, and AstraDB following MLOps best practices; integrated RAG pipelines and image processing for context-aware medical responses.

Weather Risk Intelligence Platform | *Lumen/Laravel, MySQL, HERE Maps API, NOAA SWDI*

- Built a weather intelligence engine aggregating NOAA hail and storm data for real-world risk analysis; designed normalized schemas from scratch supporting one-to-many and many-to-many relationships across weather, geographic, and business datasets.
- Built Klaviyo and SendGrid pipelines for prospect data storage, targeted outreach, and notifications.

Technical Skills

AI / LLM Systems: LLM Apps (OpenAI, Claude, Amazon Bedrock, Ollama), RAG Pipelines, Vector Search (pgvector, AstraDB), LangChain, Agentic Workflows, Streaming (SSE), Speech-to-Text, Prompt Versioning, LLM Evaluation
Backend & APIs: Node.js, NestJS, Express, Python (Flask), Laravel/Lumen, REST APIs, OpenAPI, Prisma, Better Auth, JWT
Frontend: React 19, Next.js, TypeScript, TanStack Query and Router, Vite, Tailwind CSS, PWA, i18n and Arabic RTL, Svelte
Cloud & DevOps: AWS (S3, CloudFront, UAE region me-central-1), GCP, Docker, Kubernetes, Turborepo, Git, GitHub
Security: RBAC, HttpOnly Cookie Sessions, Content Security Policy, TOTP 2FA, AES-256 Field Encryption, Presigned URLs
Languages: TypeScript, JavaScript, Python, SQL, PHP | **Databases:** PostgreSQL, MySQL, MongoDB, Redis, AstraDB

Education

Bachelor of Engineering in Computer Science and Electronics | *University of Mumbai, Mumbai* **Jul 2021 – Jun 2024**