SUBHRAJEET SWAIN

AI Engineer | Forward Deployed Engineer

Noida, India | +91 94397 00790 | subhrajeet.swain99@gmail.com | linkedin.com/in/subhrajeetswain

SUMMARY

AI engineer, 3.5 years building production software, the last two years on LLM systems - RAG over vector search, LLM parsing and scoring pipelines, and agentic workflows with tool calling, state and guardrails - owning the backend, retrieval layer, evaluation and deployment on each. On four client products I was the direct technical contact: running discovery, learning the domain from the customer, and turning an undefined ask into a working system. Most useful where the requirement is still ambiguous and someone has to sit with the client and build it.

TECHNICAL SKILLS

AI & LLM Engineering: LLM application design, RAG pipelines (chunking, embeddings, retrieval, reranking), agentic workflows and tool calling, LangGraph, LangChain, prompt engineering, output guardrails and validation, evaluation of grounding and response quality, OpenAI API, Anthropic, Google Gemini, Groq, Hugging Face Transformers, Replicate, MCP servers, TensorFlow.js in-browser inference

Vector & Data: MongoDB Atlas Vector Search, embeddings and cosine-similarity ranking, semantic search, PostgreSQL, MongoDB, Redis, Supabase

Backend: Python, Django REST Framework, Node.js, TypeScript, Express.js, REST API design, GraphQL, WebSockets, Kafka, system design, authentication and authorisation, multi-tenant architecture

Deployment: Docker, GitHub Actions CI/CD, Nginx, AWS, GCP, Vercel, environment and secret management

Frontend: React, Next.js, TypeScript, Tailwind CSS

Forward Deployed / Delivery: Direct client discovery and requirement gathering, domain modelling from scratch, technical scoping and estimation, solution demos, change-request handling, stakeholder communication across timezones, mentoring and code review

EXPERIENCE

WitArist IT Services, Noida | Jun 2025 - Present
Software Engineer Executive

- Design and ship the AI capability across the company's SaaS products: RAG pipelines over vector search, LLM parsing and scoring, agentic workflows with tool calling and conversation memory, and the guardrails that decide when the model is allowed to answer at all.
- Act as the forward-deployed technical contact for clients - run discovery and requirement calls, scope open-ended asks into deliverables, model unfamiliar business domains from those calls, demo releases and handle change requests.
- Own the backend architecture behind those products: data modelling, authentication and authorisation, multi-tenant isolation and third-party integrations.
- Run the deployment side end to end - Docker, GitHub Actions pipelines, Nginx, secret management, database migrations and staging-to-production releases on DigitalOcean, Vercel and VPS.
- Lead delivery across a team of 3 developers: breaking requirements into scope, allocating work, reviewing pull requests and setting code and documentation conventions.

APIDOTS Private Limited, Chandigarh | Apr 2023 - Jun 2025
Full Stack Developer

- Integrated LLM and vector search capability into client products, including RAG pipelines and multiple model-provider APIs.
- Was the direct client contact on assigned projects - requirement calls, scoping, release demos and change requests.
- Built and maintained REST APIs in Node.js and Express across multiple client products, covering caching, role-based access and tenant isolation.
- Worked on real-time data services with Kafka and WebSockets, and split monolithic services into Dockerised microservices behind load balancing.
- Handled containerisation and deployment: Docker, Nginx, GitHub Actions CI/CD, blue-green deploys, and infrastructure on AWS and GCP.

AI PROJECTS

AI Assessment & Proctoring Platform | Multi-tenant technical hiring SaaS | assessment-frontend-git-development-witarist.vercel.app

- Candidates take skill-matched MCQ sets, a live coding round and a scored video intro. Violations are counted server-side with weighted thresholds for warning, auto-submit and hard termination, fed by TensorFlow COCO-SSD running in the browser.
- Kept AI proctoring cheap by design: ML flags carry a confidence score and video timestamp, and Gemini only reviews a short window around high-confidence flags instead of the whole recording - a deliberate choice about where the model earns its cost.

Stack: TensorFlow.js, OpenAI, Gemini, Express, TypeScript, MongoDB, Redis, Next.js, Piston, DigitalOcean

ProTrader | LLM parsing pipeline for live financial orders

- Turns unstructured Telegram trade calls into structured broker orders with a Groq LLM and places live trades through MetaApi.

- Real money moves through this, so the parser refuses anything it cannot fully resolve rather than guessing at a partial signal - the guardrail is the product.
- Validated the parser against a hand-scored set of real Telegram calls, tracking refusal correctness over parse accuracy - a wrong refusal costs one trade, a wrong parse costs money - then tuned thresholds from production review of live orders.

Stack: Groq LLM, Node.js, Express, Prisma, MongoDB, Telegraf, MetaApi, Docker

StyleAI | AI video generation platform | stylai.fr

- Users upload images, generate video through Replicate and Kling, then compose clips with music and transitions in a multi-track editor with real-time preview.
- Server-side composition on FFmpeg with Shotstack export, shipped on Docker with GitHub Actions CI/CD.

Stack: Replicate, Kling, Next.js, Node.js, FFmpeg, Shotstack, Stripe, Firebase, Docker

Hire@Witarist | LLM intent extraction over a developer catalog | witarist.com

- Searchable catalog of 1,000+ pre-vetted developers where OpenAI handles intent extraction on the natural-language search query.
- Rebuilt as a fully pre-rendered site on Vercel CDN with an offline export, removing the runtime database entirely from a read-heavy workload.

Stack: OpenAI, Next.js, TypeScript, Redux, TanStack Query, Vercel Serverless Functions

FORWARD DEPLOYED WORK

Gold Megha | Jewellery manufacturing ERP built from the shop floor up | goldmegha.com

- I had no background in jewellery manufacturing. I learned the trade from calls with the floor - karigar ledgers, melting purity, tree numbering, WIP and metal-loss reconciliation - and named every state in the software after what the department actually calls it, so the people using it never had to translate.
- Two-sided party ledger that nets on fine weight with financial-year windows and brought-forward balances.
- Access control modelled as Supabase Row-Level Security policies rather than application-layer checks, so the rules live next to the data.

Stack: React, Vite, TypeScript, shadcn/ui, Supabase (PostgreSQL with RLS), Vitest, Playwright

Providocs | Care documentation platform for US HCBS agencies | providocs.com

- The client could not fully articulate the Ohio DODD regulatory rules; I worked them out with him on calls and built a system that has to survive a compliance audit, not just a demo.
- Multi-tenant platform for home care and community-based service agencies covering daily documentation, transportation scheduling, attendance and incident management under role-based access control.

Stack: Express, TypeScript, MongoDB, Redis, Next.js 15, Redux Toolkit, TanStack Query

BACSWN | Real-time flight telemetry and carbon emissions tracking

- Kafka and WebSocket services handling gigabytes of flight telemetry a day for live position updates, split into Dockerised microservices behind a load balancer.
- Computed per-flight carbon emissions and optimised routes from weather data and aircraft specifications, including a formula for CO2 emitted crossing a country's airspace.

Stack: Node.js, Python, Kafka, WebSockets, MongoDB, Docker

OPEN SOURCE

node-folder-structure | NPM package | www.npmjs.com/package/node-folder-structure

- Published package that scaffolds a Node.js application in a three-layer architecture with database connections, CRUD routes and server setup already wired.

EDUCATION

B.Tech, Electrical Engineering (7.8 CGPA)

Parala Maharaja Engineering College, Berhampur | 2019 - 2023