

Siddhant Vishnu

Pune, India | US citizen | +91 95455 72005 | siddhant.m.vishnu@gmail.com | [in](#) [LinkedIn](#) [X](#) [GitHub](#)

EXPERIENCE

Froncort.AI

Mar 2025 – Sep 2026

Intern → Lead Engineer (Jun 2025) → Product Manager (Jan 2026)

Pune, India

Regulatory documentation AI for drug and device submissions (IND, NDA, BLA, CER, BER, CSR)

- **Client impact:** Forward-deployed engineer for **4 pharma and medtech clients in Europe and the US**: one in production, three on paid POCs. **Ran the production deployment end to end:** releases, uptime and issue resolution. Stood up **GDPR-compliant instances** and **HIPAA-compliant instances**. Closed 66 client review comments in 24 hours.
- **Infrastructure:** Wrote reusable Terraform that makes every client environment **repeatable and isolated** (EC2, S3, ECR, Lambda, RDS, SQS, KMS), with GitHub Actions OIDC CI/CD and **full test coverage**.
- Built the Node.js (Express, Prisma) orchestrator that routes jobs over **AWS SQS** to Python ingestion, generation and PDF services.
- **AI pipeline:** Added **prompt-injection firewalls** on uploaded documents at ingestion and on users' regeneration prompts.
- Designed the **RAG** ingestion and generation pipeline (Python, Pinecone, DynamoDB), from HTML demo to production, and led its contextual retrieval. Every claim **cites its exact source excerpt**, and in-house regulatory experts confirmed the citations.
- Built **FDA- and EU MDR-aligned domain agents on adaptive runtime prompts** that assemble the context for each report type, replacing static templates with drug- and device-specific generation. A redesign made generation **~70% faster**.
- Each ~600-page report costs **under \$30** to generate, and drafting went from **weeks by hand to under 20 minutes with AI**. Also built a feature that recommends a product's FDA approval route (Gemini on Vertex AI).
- **AI metrics and evals:** Defined the **confidence and completeness scores** for every generated report on a Reliability, Alignment and Provenance layer formulated with regulatory experts, with per-section explanations.
- Co-led **Audit Mode**. Each section shows its regulatory context, evidence and reasoning, and audits flag missing requirements and submission blockers before a client files.
- Led an eval loop based on Karpathy's **AutoResearch** that scores each experiment on **regulatory acceptability, contextual precision, report faithfulness and context relevance**. Adversarial checks reject metric gaming, and only improvements are kept. Prompts it tuned consistently score **above 0.85**, and reports clear expert review in **one or two passes**.
- **Leadership:** Technical lead for **3 AI products and up to 9 engineers**. Ran daily standups across infra, full stack and AI.
- Conceived the runtime-prompt and scoring designs, handed each to an engineer with architecture plans and reviewed the builds. Led client technical discussions and **directed PRs across AI, backend, DevOps and frontend**.

AI assistant for a consumer biomarker-testing platform

- Built an AI Concierge that answers each user's personalized questions about their own biomarkers (FastAPI, LangGraph, Gemini on Vertex AI / Claude, Qdrant). It streams over SSE at **<4s time to first token** via prompt caching.
- Ran guardrails and the router in parallel to keep latency low. Queries use **user-scoped SQL templates**, so the model never writes SQL. **CI evals** gate accuracy, privacy and safety.

AI NETRA: TB outbreak prediction platform with FIND, supported by Google.org

- Architected, with a teammate, the platform's **analytics agent** (LangGraph): it ingests government Excel files in any format and answers each user query by writing and running custom Python (Pandas) in an isolated **Docker sandbox**. Every number traces to its source cell. **Demo in 2 days, production in 3 weeks**; adopted as the platform's third module.
- Reviewed all of the team's technical build plans, including the geospatial TB prediction model, and **led its field validation** with government TB officials: **55.2% less screening effort, 54.8% lower cost**. Showcased at India AI Impact Summit 2026.

Mirage

May 2026 – Present

Co-founder and engineer (team of 3). A Chrome extension that edits any website from plain English.

Pune, India

- Built the **agent pipeline** (planner, code generator, verifier, repair). A run counts as failed unless the verifier proves the change landed. **About 100 users**, 400-person waitlist, **16VC Summer Founder Fellowship**.
- Built an **LLM-as-judge release gate**: 10 held-out cases in **7 categories** (functional, visual, fused, refinement, local state, lifecycle, site-wide), each run as repeated judged trials, with criteria hidden from the agent. A release needs 90% overall, 85% per category, **zero false successes** and zero scope violations. Trials log grounding, repairs, verifier agreement, latency and cost.
- Created a Codex and Claude Code dev loop that opens Chrome, applies changes, runs tests and diffs screenshots, repeating until they match the design.
- Shipped the backend on Vercel and Supabase with JWT auth, per-token billing and OpenAI, Anthropic and Gemini adapters.

EDUCATION & AWARDS

Pune Institute of Computer Technology

Pune, India

BE Computer Engineering, CGPA 9.467/10

Class of 2027

- 1st place at IIT Bombay TechFest AI Quest 2024 (MercuryAI, multi-source RAG), 2nd place at Cummins IndoVateAI 2025, 3 hackathon wins in total. Selected for **YC Startup School Bangalore 2026**.

SKILLS

Languages and backend: Python, TypeScript, SQL, FastAPI, Node.js, Express, Prisma, React, PostgreSQL, Redis

AI: LangGraph, RAG, evals, LLM-as-judge, prompt caching, Qdrant, Pinecone; OpenAI, Anthropic and Gemini (Vertex AI) APIs

Cloud: AWS (ECS, SQS, RDS, Lambda, KMS, DynamoDB), Terraform, Docker, GitHub Actions, Vercel, Supabase