Sanidhya Ravi

Generative AI Developer

Mumbai, India - ravisanidhya@gmail.com - +91 6392173757
github.com/sanidhya1910 - linkedin.com/in/sanidhya19

SUMMARY

Generative AI developer finishing an MCA. I fine-tune small models until they beat the prompted baseline on rule-based evals, and build the serving infrastructure around them - serverless GPU pipelines, inference APIs, and the products on top. Currently AI Architect Lead Intern at Namaah Tech, where I am the top contributor on the platform repository.

EXPERIENCE

AI Architect Lead Intern

03/2026 - Present

Namaah Tech Pvt. Ltd.

Bangalore (Remote)

- Built the platform's generation canvas - episodes as board objects with stacked scenes and shots, an agent given real tools and inputs, and run state durable enough to survive a page reload.
- Closed a set of security holes across the API surface: admin takeover, SSRF, captcha bypass, and cross-user reads, plus hardening startup to fail loudly when Turnstile has no production secret.
- Designed and shipped multi-tenancy - teams, roles, context switching, trusted devices, and a per-member spend ledger surfaced in both the admin panel and the user profile.
- Integrated fal.ai inference: referencing every input by name, routing voiced shots through ref-to-video, and holding one character to one face and one voice across the whole platform.
- Top contributor on the platform repository - 85 of 428 commits across 28 contributors.
- Coordinating a 14-member cross-functional team (3 backend, 4 AI pipeline, 5 frontend, 1 Product Ops, 1 UI/UX), running sprint planning, onboarding, and technical briefings with founders and leadership.

Freelance Web Developer

2024 - Present

Independent

Remote

- Dive Azure - shipped a scuba school site to production on diveazure.in: Supabase auth with Google sign-in, a role-gated admin dashboard, and bespoke imagery generated through ComfyUI workflows rather than stock photography.
- BlueBelong - built a full-stack diving school platform covering 19 SSI course listings, bookings, a medical-clearance questionnaire, a marine-life encyclopedia, and an admin dashboard, with PWA offline support on Cloudflare Pages, Workers, and D1.
- Owned the whole path each time: information architecture, UI, backend, deployment, domain and DNS setup, and handover.

SELECTED PROJECTS

Northwind Support LoRA

- A fine-tune that replaces a 1,000-token system prompt

- Cut the system prompt from 1,079 to 99 tokens (91%) while beating the prompted baseline on every accuracy metric
- Intent accuracy 64.0% -> 98.5%; policy-citation F1 58.6 -> 90.5; escalation F1 70.0 -> 90.3
- Escalation recall 58.3% -> 87.5% - the metric that matters when a case involves injury or fraud

Qwen3-4B, PyTorch, PEFT / LoRA, Transformers, Gradio, Hugging Face

Mise

- A constrained recipe assistant, measured by rule rather than by an LLM judge

- Scaling accuracy 46.0% -> 100.0% - arithmetic verified exactly, not eyeballed
- Ingredient-use error 21.4% -> 2.8%, beating the best retrieval baseline (25.8%)
- Declines correctly 74.5% -> 100.0% of the time

PyTorch, PEFT / LoRA, Transformers, Python, Rule-based evals

Doc Agent

- Multiline OCR that returns structured data, not just text

- Handles multiline handwritten and printed input, returning schema-shaped output
- Routes extraction by requested output format (summary, table, spreadsheet)
- Deployed to Hugging Face Spaces with a GitHub Actions sync pipeline

TrOCR, Transformers, PyTorch, Python, Gradio, Hugging Face Spaces

ComfyUI Serverless Worker

- ComfyUI workflows as a serverless GPU API on RunPod

- Submits arbitrary ComfyUI API-format workflows to a serverless GPU endpoint
- Delivers output to AWS S3 or Cloudflare R2 buckets, or inline as base64
- Reworked Dockerfile and dependency set to support additional ComfyUI custom nodes

Python, RunPod Serverless, ComfyUI, Docker, WebSockets, AWS S3 / Cloudflare R2, fal.ai

TripTactix *- AI travel planning with a real price-prediction model*

- Gradient-boosting fare model (R2 ~ 0.97) trained on 600K+ Indian flight records predicts the best time to book
- LLM itinerary generation with multi-modal flight and train comparison via SerpAPI
- Full edge deployment: Cloudflare Pages Functions, D1 database, Prisma ORM

Next.js 15, TypeScript, Cloudflare Pages, Cloudflare D1, Prisma, Groq, Gemini API, Tailwind CSS

SKILLS

Languages: Python, TypeScript, JavaScript, Java, C++
Generative AI & ML: PyTorch, Transformers, PEFT / LoRA, Hugging Face, LangChain, ComfyUI, Gradio, scikit-learn
Backend & APIs: FastAPI, Node.js, Prisma, REST / WebSockets
Frontend: React, Next.js, Tailwind CSS, Framer Motion
Databases: PostgreSQL, SQL, MongoDB, SQLite / D1, Supabase
Infrastructure & Deployment: Docker, Cloudflare Workers, RunPod Serverless, fal.ai, AWS EC2 / S3, Vercel, Git, GitHub Actions

EDUCATION

Master of Computer Applications - 8.56 CGPA 2024 - Present
Vivekanand Education Society's Institute of Technology Mumbai, India

Bachelor of Computer Applications - 7.2 CGPA 2020 - 2023
Greater Noida Institute of Technology Greater Noida, India

CERTIFICATIONS

IBM Full Stack Software Developer (Coursera) | AWS Cloud Solutions Architect (Coursera) | SQL: Querying Databases (Coursera) | MongoDB Developer (MongoDB)