

NITHIN N

AI Engineer | Agentic AI, LLM Systems & RL Environments

nithin958595@gmail.com · 6374702575 · Mysore, Karnataka · <https://github.com/oxnithinmys>,
<https://linkedin.com/in/nithin-nagaraj-a4b837275/>, https://leetcode.com/u/Nithin_VIT

SUMMARY

Integrated AI student and AI engineer with over a year of hands-on experience across four internships, building LLM applications, RAG and document-processing pipelines, and multi-agent systems shipped as FastAPI services on Docker and GCP. Building RL environments for agents; 10+ merged pull requests to Mesa. Architected a 4-agent LangGraph system with hybrid RAG that cut hallucinations ~30% and retrieval errors ~35% in internal benchmarks; fine-tuned a ViT classifier for a 30% accuracy gain. IBM AI Engineering certified; winner of a national-level blockchain hackathon (Alkimi Exchange x HackerEarth FWD 2025) and Google Agentic AI Day Hackathon 2025 finalist (Bengaluru).

EXPERIENCE

Divan Technologies 2025 – Present
Generative AI Engineer Intern Bengaluru

- Build backend AI agents for client projects with LangChain and OpenAI APIs, automating multi-step workflows through agentic decision-making pipelines.
- Developed and deployed a company website chatbot with LLM-powered conversational flows, RAG and custom prompt libraries for accurate, context-aware responses.
- Own tickets end to end (design, build, test, review) in Python and FastAPI, working with cross-functional teams.

Technologies: LangChain, OpenAI APIs, Python, FastAPI

IIT Indore 2025-01 – Present
Research Intern, Computer Vision & Agentic AI Remote

- Fine-tuned a ViT image classifier on a domain-specific dataset, improving accuracy by 30% over baseline through attention-mechanism tuning and batch optimization.
- Built a hybrid agentic pipeline combining ViT visual features with an LLM reasoning layer for context-aware, explainable predictions.
- Integrated LangGraph-style multi-agent orchestration that routes low-confidence predictions to a verifier agent for secondary analysis.

Technologies: ViT, LLM, LangGraph

DeepInsightX 2024
Software Developer Intern, Gen AI & Backend Remote

- Architected a generative-AI podcast production pipeline on GCP; containerized services with Docker and automated deployments via CI/CD, cutting release turnaround by ~60%.
- Built a Redis caching layer for AI-driven API endpoints, reducing response latency and raising throughput under concurrent load.
- Integrated Cloudflare CDN and WAF rules, improving delivery speed and hardening the service against DDoS and injection attacks.

Technologies: GCP, Docker, Redis, Cloudflare CDN, WAF

Bista Technologies Inc. 2024
Software Developer Intern Remote

- Refactored a Next.js + TypeScript codebase and expanded unit and integration test coverage in CI/CD, improving performance and release reliability.

Technologies: Next.js, TypeScript, CI/CD

SKILLS

Languages: Python, TypeScript, JavaScript (Node.js), C++, Bash

Frameworks: LangGraph, LangChain, FastAPI
Databases: ChromaDB, Pinecone, FAISS, Elasticsearch/BM25, PostgreSQL, MongoDB, Redis
Cloud & DevOps: Google Cloud Platform (Compute, Cloud Run), AWS, Azure, Docker, GitHub Actions CI/CD
System Design: REST APIs, microservices, async workers (Redis RQ)
Other: OOP, Data Structures & Algorithms, LLMs, prompt engineering, hybrid RAG, multi-agent orchestration, OpenAI API, Gemini, Groq (Llama 3), Hugging Face, spaCy, Microsoft Presidio, NER, PII/PHI detection, Vision Transformers (ViT) fine-tuning, image classification, InsightFace, model evaluation and benchmarking, RL environments for agents, LLM pretraining, research-paper implementation, Pytest

EDUCATION

VIT Bhopal, Bhopal, MP	2023 – 2028
<i>Integrated [URL], Artificial Intelligence</i>	8.1

PROJECTS

PII-Backend

HIPAA-Oriented PII/PHI Detection Service

- Built a document-processing service that detects and masks PII/PHI by combining Presidio, spaCy NER and optional Gemini LLM validation, with custom recognizers for Indian ID formats (Aadhaar, PAN).
- Implemented HIPAA-oriented risk scoring across 17 entity types with bounding-box mapping; covered by a Pytest suite and containerized with Docker for cloud deployment.

Technologies: FastAPI, Microsoft Presidio, spaCy, Gemini, Docker, Pytest

Loraframe (IDLock Engine)

Persistent AI Character & Video Generation Platform

- Built an AI pipeline with episodic memory (RAG + vector DB), an LLM prompt engine (Llama 3) for context-aware scene generation, and InsightFace identity embeddings that keep faces consistent across images and video.
- Architected the async backend on FastAPI + Redis RQ workers with PostgreSQL and Pinecone/FAISS, integrating Gemini/Imagen/Veo for multi-modal generation; added a self-healing "Refiner" worker that detects and corrects identity drift automatically.

Technologies: FastAPI, Redis RQ, Groq (Llama 3), Gemini/Imagen/Veo, InsightFace, Pinecone, PostgreSQL, RAG, vector DB, FAISS

ACHIEVEMENTS

- Winner, National-Level Blockchain Hackathon (Alkimi Exchange x HackerEarth FWD 2025)
- Finalist, Google Agentic AI Day Hackathon (2025, Bengaluru)
- Finalist, Global Health Hackathon

