

Animesh

Applied GenAI Engineer · LLM Systems · RAG + Agentic AI

animesh09021998@gmail.com · +91-8439725103 · github.com/Animesh98 · linkedin.com/in/animesh-0902-sinha

SUMMARY

- **1.5+ years** shipping LLM-powered products, RAG pipelines, and real-time agentic AI in production at Formee Holdings
- Production **RAG** (LangChain, LlamaIndex, FAISS) · **NL-to-SQL** with safety guardrails · Real-time **agentic AI** in WebRTC + Asterisk/Twilio voice/chat CRM
- **Agentic development** with **Claude Code** (1+ year daily driver) – skill-based workflows, custom slash commands, end-to-end rapid prototyping
- **IIT Bombay** graduate coursework in CS (Intelligent Systems), CPI **8.72/10** (*thesis not defended; no degree awarded*)

TECHNICAL SKILLS

LLM/GenAI: LangChain, LlamaIndex, OpenAI API, Prompt Engineering, RAG Systems, Fine-tuning (QLoRA/LoRA), Hugging Face, FAISS, Ollama, MLflow, Weights & Biases

ML/DL: PyTorch, Deep Learning, Computer Vision, OpenCV, Scikit-learn, NumPy, Pandas

Backend & Frontend: FastAPI, Flask, Next.js, React, Express.js, Azure Functions, Docker, PostgreSQL, Redis, MySQL, REST APIs, WebRTC, Tailscale, Ubuntu Server, Git

Languages: Python, JavaScript/TypeScript, C++, SQL

Agentic Development: **Claude Code** (1+ year daily driver), skill-based workflows, custom slash commands, end-to-end rapid prototyping

EXPERIENCE

Formee Holdings

R&D Software Engineer – AI/ML

Bengaluru, India
Aug 2024 – Present

- Set up a **RAG pipeline** over scraped domain corpora with vector databases to generate domain-specific academic assessments. Replaced generic LLM output with grounded, syllabus-aligned content.
- Wrote an **auto-grading plugin for Moodle** (OpenAI GPT-4.1) that grades submissions against rubrics in batch or individually, auto-triggers on deadlines, and emails evaluators. Cut manual grading by **~80%**.
- Built content generation APIs (OpenAI + LangChain) that turn course outlines into assessments, study guides, and slide decks. Added **LaTeX automation** for **100+ page** PDF compilation in a single run.
- Built a student support **CRM with Next.js + WebRTC** for real-time chat and voice. LlamaIndex-driven AI suggests agent responses from live conversation transcripts (DB-aware, with safety guardrails). Extended into a full call center using **Asterisk + Twilio** with agentic post-call lead analysis.
- Ran internal **LLM/RAG workshops** and onboarded **4 engineers** to the AI/LLM stack.

PROJECTS

- **Llama 3.1 8B QLoRA SFT → DPO Pipeline** – Built the full SFT→DPO fine-tuning loop on a homelab **RTX 4060 Ti** (**Unsloth + 4-bit NF4**) to characterize what each phase actually does to model behavior. SFT on a curated **Dolly-15k** subset, then DPO on **UltraFeedback** preference pairs. Evaluated with held-out perplexity, **Claude Sonnet** as pairwise LLM-judge with A/B position swap, mini-IFEval, a 30-prompt qualitative probe, and cross-checkpoint weight-delta diagnostics. **MLflow + Weights & Biases** tracking.
- **Jarvis Dashboard** (*Personal, built with Claude Code*) – Homelab control panel built with **FastAPI + Next.js/React**. System monitoring, Docker container management with live logs, Jellyfin media integration, mood-based movie discovery, file explorer.
- **Homelab Infrastructure** (*Personal, built with Claude Code*) – Self-hosted homelab for media automation and local LLM experimentation – **10+ Dockerized services** (Jellyfin, Sonarr, Radarr, PostgreSQL, Redis) behind **Tailscale VPN** and UFW, with **Ollama** running local inference on an RTX 4060 Ti 16GB for offline GenAI work.
- **Neural Rendering for AR** (*MS Thesis, IIT Bombay; Adv: Prof. Parag K. Chaudhuri*) – Novel NeRF-based approach to **occlusion** and **illumination** challenges in AR scene augmentation. Extended individual NeRFs as composable blocks via a Generalizing module for intermediate merged renderings.
- **Anime Recommender System** (*CS 725, IIT Bombay*) – Recommender on **13k anime, 7.2M reviews**: collaborative filtering (memory + model-based), content-based, popularity baseline. Addressed cold start. Prof. Preethi Jyothi.

EDUCATION

Indian Institute of Technology Bombay

Graduate Coursework in Computer Science (Intelligent Systems)

Jul 2021 – Jul 2024

CPI: 8.72/10

Coursework + research projects completed; **thesis not defended; degree not awarded**.

MS Thesis: Neural Rendering for AR (Adv: Prof. Parag K. Chaudhuri). Teaching Assistant for 4 CS courses including Computer Vision, Computer Graphics, and AI Programming.

Gurukula Kangri Vishwavidyalaya, Haridwar – B.Tech CSE

2016 – 2020 · **CPI: 8.63/10**

ACHIEVEMENTS

GATE CSE 2021: 99.56 percentile (100,000+ candidates) · Poster at **CSE Research Symposium 2023**, IIT Bombay (Neural Rendering for AR)