SANSKAR NANEGAONKAR

AI / ML Engineer
Python · PyTorch · LangGraph · RAG · MCP · Evals · Fine-tuning · Multimodal · Production Systems
sanskarnanegaonkar@gmail.com | linkedin.com/in/sanskar-nanegaonkar | github.com/02san02 | sanskar-nanegaonkar.com
India | Transferable H1B, open to US relocation

PROFESSIONAL SUMMARY

AI engineer with **4+ years** shipping production LLM, agentic, multimodal, and clinical ML systems end-to-end: enterprise Retrieval-Augmented Generation (RAG) for **NTT DATA** and two more customers, voice agents across **1,000+** calls, and the clinical model behind an **FDA Breakthrough Device Designation application**.

EXPERIENCE

I'mbesideyou Inc. *(Applied AI startup)* Tokyo, Japan (Remote)

Senior AI Engineer, Agentic AI & Production Systems *Jan 2025 – Present*

- Architected and shipped a multi-tenant HR-policy RAG assistant for **NTT DATA** and two enterprise customers, indexing **1,000+** HR policies: FastAPI service with **Qdrant** per-tenant retrieval, hybrid search with cross-encoder reranking to separate near-identical policies that differ by employee category or geography, and parent-document retrieval to resolve cross-page references; OpenAI GPT synthesis, safety-aware policy-conflict detection, source-priority rules, and **Langfuse** audit traces.
- Owned multimodal SOP-generation and gap-analysis modules that convert desktop activity into operating procedures: multi-provider LLM pipeline (Gemini vision, GPT synthesis, Claude critic), **LangGraph** agentic multi-stage orchestration, Pydantic structured outputs, and write-review-rewrite loops; deployed across **3+ organizations** with **100+ operators**.
- Reduced manual transcript review per agent from **a week** to **a few automated hours** using a **multi-agent RAG** optimization system: planner / extractor / synthesizer chains, RAG over source corpora, human-feedback interpreter, variant regeneration router, and regression scoring against correctness thresholds.
- Shipped voice AI agents on **Vapi** and **LiveKit** for lead qualification, interviewing, and structured data collection; LLM dialogue policy with Pydantic tool calls and transcript-grounded evals matched human-agent pickup and conversion rates across **1,000+** calls.
- Replaced random spot-check QSC audits with continuous automated **VLM** compliance coverage, piloted live across 2 restaurant locations; designed a **VLM**-assisted auto-labeling pipeline that bootstrapped a proprietary training dataset in a kitchen domain with no open-source data or model coverage.

Data Scientist, Clinical AI & Multimodal ML *Aug 2022 – Dec 2024*

- Built a multimodal depression-severity model using **DeepFace** facial emotions and **librosa** speech-prosody features (early fusion); reached **83%** accuracy and extended to bipolar-vs-unipolar classification at **78%** across **350+** patients.
- Productionized the clinical ML pipeline on **AWS**; collaborated with Keio on an **FDA Breakthrough Device Designation application** and Pre-Sub filings, and validated depression-severity predictions with Northwell Health and Hamamatsu across clinical cohorts.
- Cut therapy documentation time by **70%** with an LLM system converting transcripts into Pydantic-validated SOAP notes; a clinician style-adaptation loop distilled edits into reusable per-clinician profiles.

Data Science Intern *May 2021 – Jul 2022*

- Built video analytics pipelines across classroom sessions, live-streaming, and sales call recordings; extracted behavioral and linguistic signals (Python, OpenCV, scikit-learn) to distinguish high from low performers across **10,000+** recordings and generated structured coaching recommendations.

PROJECTS

Preference Optimization & Alignment Stack *PyTorch, TRL, LoRA / QLoRA, DPO, vLLM*

Fine-tuned **Llama 3 8B** for medical Q&A (MedQA / PubMedQA): curated instruction and preference data, ran LoRA / QLoRA domain SFT followed by DPO preference optimization, and evaluated against the base model and a frontier API on task success, helpfulness, latency, and cost before quantized vLLM serving.

LLM Evaluation Harness *PyTest, Pydantic, Langfuse, LLM-as-judge, GitHub Actions*

Built a pytest-style evaluation harness for LLM-powered services: golden-set fixtures with versioned prompts, **LLM-as-judge** scoring with confidence-aware aggregation, per-test cost and latency budgets, and a GitHub Actions gate that blocks prompt or model changes regressing curated benchmarks. Plugs into Langfuse traces for failure inspection.

SKILLS

LLM Apps: LLM orchestration, agentic systems, RAG (hybrid retrieval, cross-encoder reranking, BM25), agents (tool use, ReAct, MCP), LangGraph, prompt engineering, Pydantic structured outputs, Claude / OpenAI / Gemini APIs, multi-step pipelines, voice (Vapi, LiveKit)
Evals & Observability: LLM-as-judge, RAGAS, golden sets, regression scoring, Langfuse, LangSmith
AI / ML: PyTorch, HuggingFace Transformers, fine-tuning, PEFT, LoRA / QLoRA, vLLM, GPTQ / AWQ, multimodal fusion, VLMs, scikit-learn
Post-training / Alignment: RLHF, reward modeling, DPO, PPO, preference datasets
Backend / Infra: Python, SQL, FastAPI, PostgreSQL, Qdrant, FAISS, Docker, Kubernetes, AWS, HIPAA compliance

EDUCATION & PUBLICATIONS

Indian Institute of Technology (IIT) Hyderabad, B.Tech, Engineering Physics, CGPA **8.48 / 10** *2018 – 2022*

Nanegaonkar & Ando. *A Practical Approach to Predicting Depression: Verbal and Non-Verbal Insights with Machine Learning.* HIIJ 13(2), May 2024.
Co-author, *Chatbot Interventions for Early-Stage Depression and Anxiety: A Pilot Study* (EHPS Conf., Aug 2025).