

SAI TARUN REDDY LANKIREDDY

Software Engineer | Backend & AI Systems

Cincinnati, OH (Open to Relocation) | saitarunreddylankireddy@gmail.com
tarunlankireddy.com | linkedin.com/in/sai-tarun-reddy-lankireddy | github.com/tarunn31

PROFESSIONAL SUMMARY

Software engineer with 3+ years building Python backends, REST APIs, and LLM systems in healthcare and enterprise. Shipped RAG and agent workflows for 500+ internal users at Molina Healthcare, and built Nita, an AI wellness app, as a solo founder.

TECHNICAL SKILLS

Languages & Backend: Python, SQL, Java, JavaScript, React, FastAPI, Flask, REST APIs, Microservices, OAuth, OpenAPI, pytest
Data: PostgreSQL, MySQL, Supabase, Redis, pgvector, Pinecone, FAISS, ChromaDB, Pandas, NumPy, ETL Pipelines
LLM & Agents: OpenAI, Claude, Azure OpenAI, Amazon Bedrock, LangChain, LangGraph, MCP, RAG, Hybrid Search, Reranking
Evaluation & Safety: RAGAS, LangSmith, LLM-as-Judge, Golden Test Sets, Guardrails, Prompt-Injection Defense, PHI Redaction
ML & Cloud: PyTorch, Hugging Face, BERT, spaCy, MLflow, AWS (SageMaker, EC2, S3, Lambda), Azure, Docker, GitHub Actions

EXPERIENCE

Founder & Software Engineer | NITA CARE LLC | Cincinnati, OH Mar 2026 – Present

- Built Nita, an AI wellness app for iOS and Android, as a solo founder using React Native (Expo), a FastAPI backend on Railway, and Supabase (PostgreSQL 17, pgvector, Auth, Storage). Public launch Oct 2026.
- Wrote the pattern engine in plain Python with no ML libraries: regression trends, z-score anomaly checks, and lagged Pearson correlations that only show up after 10 paired days of data.
- Kept the AI from inventing health claims. Claude only explains numbers the engine computed, the prompt bans any number it was not given, and failed correlations are passed in so it never implies them.
- Added AI consent checks on client and server: consent is stored as versioned, append-only records, every AI endpoint returns 403 without it, and user identity always comes from the auth token.
- Cleared App Store and Google Play review with AI consent screens, health data kept off iCloud backups, and lean permissions.
- Also built chat memory (OpenAI embeddings in pgvector), voice chat, FCM reminders, and scheduled weekly digest jobs.

Systems Test & Debug Technician (Contract) | Jabil, via Lorien | Florence, KY Aug 2026 – Present

- Debug data center servers on the production test line: read test logs, run diagnostics, find the root cause, and fix the unit.
- Use AI tools to turn test logs into step-by-step fix procedures and record failure data used to train an internal AI tool.

AI Software Engineer (Contract) | Molina Healthcare | Remote Jun 2025 – Mar 2026

- Built a RAG platform over 50K+ clinical and policy documents with semantic chunking, metadata enrichment, and Pinecone and pgvector retrieval, cutting clinician document lookup time 40%.
- Worked with clinical SMEs and product managers on LangGraph agent workflows that plan, call tools, and summarize authorization packets, reducing manual review effort 35%.
- Lowered measured hallucination rates 30% across OpenAI and Claude models using function calling, structured outputs, and checks that answers are grounded in retrieved text.
- Set up RAGAS and LangSmith evaluations over 200+ curated test cases and made them a required gate for every release.
- Wrote FastAPI microservices and REST APIs for LLM inference and retrieval, used by 500+ internal users across 4 workflows.
- Added PHI redaction, token auth, and audit logging for HIPAA, containerized services with Docker, and set up GitHub Actions CI/CD with latency and token monitoring. Releases went from biweekly to weekly.

Associate Software Engineer | NTT DATA | Hyderabad, India Jan 2022 – Aug 2024

- Wrote Python backend services and REST APIs for 5+ enterprise apps, standardizing data validation and ML inference.
- Built SQL and Pandas ETL pipelines over 1M+ records, plus BERT and spaCy NLP for document classification and entity extraction.
- Deployed models to AWS SageMaker endpoints and FastAPI services, tracking experiments and model versions in MLflow.

PROJECTS

AI Data Analyst Copilot | Python, OpenAI GPT-4.1, Streamlit, Pandas, Plotly GitHub | Live Demo

- Built a Streamlit app where GPT-4.1 turns plain-English questions about a CSV into pandas code run in a restricted sandbox.
- Failed code returns to the model with its error for up to 3 fixes; it also auto-builds KPI dashboards and data-quality reports.

Sign Language Translator | JavaScript, deeplearn.js, KNN, Web Speech API Live Demo | Paper

- Built most of the code and led testing for a team app that learns a new sign from 30 webcam samples and translates it live.

EDUCATION, PUBLICATION & CERTIFICATIONS

M.Eng., Computer Science | University of Cincinnati, Cincinnati, OH | GPA 3.8/4.0 May 2026

B.Tech., Information Technology | Prasad V. Potluri Siddhartha Institute of Technology (JNTUK), India | GPA 3.7/4.0 Apr 2024

Publication: Co-author, "Dynamic Sign Language Detection," Grenze Int. Journal of Engineering and Technology, 2025 (Scopus)

Certifications: IBM Generative AI Engineering Professional Certificate | ServiceNow Certified System Administrator