

SAHIL KUMAR NAYAK

+91-9348881839 sahilsony2002@gmail.com linkedin.com/in/sahil-kumar-nayak github.com/Sahil130202

Professional Summary

Applied AI & Backend Engineer specializing in taking Generative AI systems from prototype to production. Proven ability to architect local-first, privacy-preserving platforms using asynchronous Python (FastAPI), PostgreSQL (pgvector), and decoupled microservices (Celery/Redis). Bridges the gap between complex LLM workflows (RAG, Agents) and robust, highly scalable backend infrastructure to deliver performant, low-latency applications.

Technical Skills

Backend & Systems: Python 3.11+ (Asyncio), C/C++, FastAPI, SQLAlchemy 2.0, Celery, Redis, Server-Sent Events (SSE)
Generative AI & LLMs: RAG Pipelines, Agentic Workflows, LangChain, OpenAI API, Ollama, Hugging Face
Databases & Vector Search: PostgreSQL, pgvector, ChromaDB, HNSW Similarity Search, Alembic
Cloud, DevOps & Full-Stack: Docker, Docker Compose, AWS, Git, Next.js 16 (App Router), TypeScript, Zustand
ML & Data Tooling: TensorFlow, PyTorch, Scikit-learn, Pandas, MLflow

Experience

Labmentix Technologies

Jun 2025 – Dec 2025

AI & Backend Engineering Intern

Remote

- Architected automated data processing pipelines in Python to ingest and transform large-scale datasets, reducing ML model iteration and training time by **40%**.
- Packaged predictive models (TensorFlow/PyTorch, Scikit-learn) into robust REST APIs, ensuring high availability, reliability, and sub-second inference latency.
- Established automated benchmarking and evaluation workflows to continuously track system performance, data throughput, and model accuracy across multiple real-world deployments.

IIT Chennai — E&ICT / Eduxlab

2022 – 2023

Software & ML Research Intern

Chennai, India

- Engineered memory-efficient data processing pipelines using TensorFlow/Keras and Rasterio for CNN-based satellite imagery analysis (E&ICT, 2023).
- Built classification and clustering algorithms using Python (Pandas, Scikit-learn) to extract actionable insights from large-scale tabular datasets (Eduxlab, 2022).

Engineering Projects

Cortex AI OS | Next.js, FastAPI, PostgreSQL, Redis, Celery, Ollama, Docker | [GitHub](#)

2026

- Architected a 6-service microservices monorepo deployed via a custom Docker Compose bridge network, guaranteeing **100% local data privacy**.
- Engineered a concurrent API layer with FastAPI, Pydantic v2, and asyncpg, exposing 13 REST endpoints secured end-to-end with short-lived JWT access and HTTP-only refresh tokens.
- Designed an advanced RAG ingestion pipeline backed by PostgreSQL (pgvector) indexing **1536-dimensional** embeddings optimized for HNSW semantic search.
- Implemented SSE infrastructure for real-time LLM response streaming to the Next.js client, ensuring **sub-100ms** perceived latency.

AI Tutor Bot | LangChain, GPT-4, ChromaDB, Python | [GitHub](#)

2025

- Developed an end-to-end conversational agent using LangChain and GPT-4 to ingest, chunk, and synthesize complex PDF textbooks into context-grounded responses.
- Engineered multi-turn stateful memory (ConversationBufferWindowMemory, k=5) and integrated DuckDuckGo Search as a live fallback layer to mitigate hallucinations.

Custom UNIX-Like Shell | C/C++, POSIX System Calls

2024

- Developed a from-scratch shell in C/C++ using low-level POSIX calls (fork, execvp, waitpid) with support for foreground/background execution and robust command parsing.

Research & Publications

Real-Time Device-Free Human Activity Recognition Using WiFi CSI and Deep Learning

Accepted 2026

AIMLCPS 2026

- Designed a hybrid CNN-GRU deep learning architecture to process high-dimensional, noisy WiFi Channel State Information (CSI) for real-time spatial inference.

Education

University of Hyderabad

2024 – 2026

M.Tech, Artificial Intelligence

Hyderabad, India

VSSUT, Burla

2019 – 2023

B.Tech, Artificial Intelligence

Burla, Odisha