AAYUSH SAHU

Applied AI Engineer · LLM Systems & RAG · Founder @ OctiqAI · Data Analytics @ Suzlon
aayushsahu0406@gmail.com · Pune, Maharashtra · LinkedIn · GitHub

SUMMARY

Second-year CSE student at MIT-ADT University (AI & Edge Computing) running three concurrent roles: Founder & CEO of OctiqAI, Data Analytics Intern at Suzlon Energy, and AI Engineer at SYNC. Specialises in offline LLM systems, RAG pipelines, multi-model orchestration, and ML-driven analytics on industrial sensor data. Shipped 4 production AI products; authored an independent research paper on resource-constrained offline AI architecture; Cambridge CEFR B2 certified.

EXPERIENCE

Founder & CEO | **OctiqAI** *May 2026 – Present · Remote*

- Architected a vector-embedding organisational memory engine with sub-200ms semantic retrieval latency on local hardware — zero cloud dependency.
- Shipped core FastAPI + agentic orchestration backend within 5 weeks of founding; building toward public beta.
- Designed multi-session context persistence layer that eliminates information loss across fragmented startup workflows.
- OctiqAI | Contextual Intelligence for Founders

FastAPI · ChromaDB · Anthropic API · Agentic Orchestration · Async Python

Data Analytics Intern | **Suzlon Group** *Jun 2026 – Jul 2026 · Pune, On-site*

Engineered a Python-based SCADA analytics pipeline to process **280K+ wind turbine event records**, implementing event filtering, FIFO pairing, T-10 normalization, duration calculation, and automated Excel report generation.

Designed and developed an **AI-powered Document Intelligence platform** for extracting structured text, tables, images, and flowcharts from engineering PDFs using deterministic parsing and LLM-assisted processing.

Built and integrated AI capabilities into an **Enterprise 8D Root Cause Analysis (RCA) platform**, including intelligent autofill, analytics dashboards, and a context-aware conversational AI assistant.

Conducted research on wind turbine systems, SCADA architecture, failure analysis, and RCA methodologies; additionally served as **Site Visit Coordinator** for the internship cohort, coordinating industrial site visits and communication.

Python · Scikit-learn · Random Forest · Isolation Forest · Pandas · Time-Series Analysis

Artificial Intelligence Engineer | **SYNC — Internship** *Jan 2026 –Jun 2026· USA, Remote*

- Built multi-model LLM orchestration layer routing tasks by complexity — reduced average inference cost per query by ~35%.
- Developed production RAG pipelines with semantic chunking; reduced hallucination rate across enterprise test sets.
- Architected async AI infrastructure processing 500+ daily automated workflow actions.

FastAPI · RAG · Multi-Model Orchestration · Vector Databases · Async Python

Web Content Writer | **UdaanScholars — Internship** *Apr 2026 – Present · Remote*

- Published 18+ long-form college profile articles (avg. 1,800 words); contributed to 22% uplift in organic search visibility within 6 weeks.
- Built structured content templates adopted team-wide, reducing per-article production time by ~25%.

SEO Writing · Google E-E-A-T · Content Strategy

Management Team Member & Social Media Coordinator | **ACES MITSOE** *Aug 2025 – Apr 2026 · Pune, On-site*

- Executed 7+ technical events with 800+ combined attendees; grew ACES social media engagement 38% over 6 months.

Founder | **Luné Aesthetics** *Feb 2022 – Dec 2024 · India*

- Founded and scaled a student accessories brand to 500+ units sold; managed full operations — sourcing, pricing, branding, and sales — across 3 years.

KEY PROJECTS

E.D.I.T.H. V8 — Offline Personal AI Assistant *Jun 2025 – May 2026*

- Built a fully offline, multi-model AI assistant (Llama 3.2 / Qwen 2.5 / CodeLlama / LLaVA) with intelligent task-based LLM routing on WSL2.
- Integrated ChromaDB RAG memory, Whisper STT, Piper/Kokoro TTS, Telegram remote control, ADB Android automation, and APScheduler — zero cloud dependency.
- Documented system architecture in an independent research paper on resource-constrained offline AI design.

Python · FastAPI · WebSocket · Ollama · ChromaDB · Whisper · Piper TTS · React

Maneuver — Real-Time Voice AI Assistant *May 2026*

- Built real-time voice AI using LiveKit WebRTC + Groq Llama 3.3 with sub-second response latency; synchronised live transcription in React frontend.

LiveKit · WebRTC · Groq API · Whisper STT · Edge-TTS · React · Python

Reel Craft AI — AI Caption Generator *Apr 2026*

– Designed and deployed a Claude API-powered caption tool for Instagram Reels; achieved 286+ LinkedIn impressions within 24 hours of launch.

React · Anthropic Claude API · Vercel · Prompt Engineering

SKILLS

AI & LLMs: RAG, LLM Orchestration, Agentic Workflows, Prompt Engineering, Ollama, ChromaDB, Pinecone, Anthropic API, LLMOps
ML & Analytics: Random Forest, Isolation Forest, Scikit-learn, Predictive Maintenance, Time-Series Analysis, MLflow, Pandas, NumPy
Backend: Python, FastAPI, WebSocket, REST APIs, AsyncIO, PostgreSQL, SQLite, TimescaleDB
Frontend: React, JavaScript, TypeScript, HTML/CSS, Tailwind, Next.js, Vite, Vercel
Voice & Realtime: Whisper STT, LiveKit, WebRTC, Edge-TTS, Piper TTS
Tools: Git, GitHub, Docker, WSL2, ADB, Telegram Bot API, Figma

EDUCATION

B.Tech — Computer Science & Engineering (AI & Edge Computing) *Aug 2025 – May 2029 (Expected)*
MIT-ADT University, Pune · Core Committee Member, ACES

CERTIFICATIONS & RESEARCH

Foundations in Generative AI — IBM · *Apr 2026* · ID: ALM-COURSE_3955079
Introduction to Prompt Engineering — Simplilearn · *Apr 2026* · ID: 10119751
Linguaskill English Proficiency (CEFR B2) — Cambridge University Press & Assessment · *Mar 2026* · ID: L0000024134
Python Training (Spoken Tutorial) — IIT Bombay · *Apr 2026*
C Programming Training — IIT Bombay · *Nov 2025*

Research Paper: "EDITH: A Resource-Constrained Offline Personal AI Architecture with Multi-Stage Routing and Hybrid Memory on Consumer Hardware"
github.com/Aayushashsahu/EDITH-