

# OJAS SONI

+91-7696968251 | [ojassoni0221@gmail.com](mailto:ojassoni0221@gmail.com) | [LinkedIn](#) | [GitHub](#) | [Personal Blog](#)

## SKILLS

---

**Languages:** Python, C/C++, SQL

**GenAI & LLM:** LLMs, Retrieval-Augmented Generation (RAG), Agentic AI, Multi-Agent Systems, Prompt Engineering, Fine-Tuning (QLoRA/PEFT), Model Evaluation, MCP

**Frameworks:** LangGraph, LangChain, CrewAI, Hugging Face Transformers, Sentence Transformers, vLLM, TensorFlow, PyTorch, FastAPI

**ML & NLP:** Machine Learning, Deep Learning, Natural Language Processing (NLP), Semantic Search, Embeddings, Reranking

**Cloud & MLOps:** AWS (EC2, S3), Docker, CI/CD, GitHub Actions, LLMOps, REST APIs, Git, Linux

**Data:** MongoDB, PostgreSQL, Kafka, FAISS

## EXPERIENCE

---

**Tata Consultancy Services** | GenAI Engineer

Noida, India

Client: **APPLE**

Nov 2024 – Present

### AI-Driven Automated Ticket Resolution System

- Engineered an **agentic incident automation platform** integrating MCP servers with a **real-time Kafka ingestion pipeline** processing **10,000+ production tickets daily** and built a RAG-powered orchestration agents to autonomously analyze and resolve production incidents.
- Developed specialized AI agents for **Splunk log intelligence**, anomaly detection, and contextual error summarization, integrating **enterprise ITSM ticketing portals** for real-time incident ingestion and tracking, and built **FastAPI microservices** with async endpoints for agent orchestration, health monitoring, and audit logging, containerized via **Docker** on scalable cloud infrastructure.
- Implemented a **RAG-driven root cause analysis pipeline** leveraging historical incident embeddings, semantic retrieval, and structured remediation generation, reducing **manual triage effort by 40%** and average resolution time from hours to minutes.

### AI-Powered Talent Evaluation and Interview Intelligence System

- Built an **agentic AI resume screening and interview automation platform** using CrewAI and Gemini, with autonomous agents for **resume parsing, dynamic question generation, and skill-gap analysis**, leveraging RAG-powered retrieval of JD-specific questions from **MongoDB** to enable end-to-end candidate evaluation.
- Implemented a **multi-dimensional evaluation framework** scoring candidates on correctness, clarity, depth, and relevance using **Sentence Transformers** for semantic similarity, achieving **87% agreement** with human interviewer decisions.
- Built robust session management with **FastAPI** handling **100+ concurrent interviews**, MongoDB persistence for conversation history, and **sub-2s response latency** across all agent interactions.
- Developed an **AI-powered summary generation agent** producing structured hiring reports (strengths, skill gaps, risk factors, recommendations) with automated **PDF generation via ReportLab**, reducing end-to-end interview processing time by **60%**.

## PROJECTS

---

### Enterprise Secure Multi-Tenant RAG Assistant

- Architected a secure **multi-tenant Agentic RAG platform** with **RBAC-driven document-level authorization**, metadata-aware retrieval, and enterprise-grade knowledge isolation across dynamic document repositories.
- Engineered scalable **LLMOps pipelines** for automated document ingestion, chunking, embedding orchestration, and **hybrid semantic retrieval** leveraging VectorDBs, reranking, and real-time synchronization workflows.
- Integrated **open-source LLMs** with citation-grounded generation, conversational memory, audit telemetry, and high-throughput **FastAPI-based inference orchestration** for production-scale enterprise AI assistants.

## EDUCATION

---

**Thapar Institute of Engineering and Technology**

*B.E. Computer Engineering, CGPA: 8.00*

**Punjab, India**

*Aug 2020 – July 2024*