Sonu Kumar

+91 7595020599 sonuhits047@gmail.com LinkedIn GitHub Portfolio

SUMMARY

AI Engineer with hands-on experience in building **LLM-based and Generative AI agents**, fine-tuning computer vision models like YOLOv8, and backend development using Java Spring Boot. Skilled in **Natural Language Processing (NLP)**, LangChain, LangGraph, OCR, RAG, and modern ML/NLP stacks with a proven ability to deliver scalable and secure solutions.

EDUCATION

Masai School — **Bengaluru, India**
Full Stack Development — *June 2023 – Nov 2023*

Dream Institute of Technology — **Kolkata, India**
B.Tech in ECE — *July 2019 – July 2022*

Holy Mary Institute of Technology and Science — **Hyderabad, India**
Diploma in EEE — *July 2014 – June 2017*

TECHNICAL SKILLS

Languages: Python, Java, JavaScript, SQL

AI/ML: Generative AI, Natural Language Processing (NLP), LangChain, LangGraph, OpenAI, Hugging Face Transformers, Scikit-learn, Pandas, Numpy, OpenCV

ML Architectures: CNN, YOLOv8, BERT, LSTM, Retrieval Augmented Generation (RAG)

Backend: Spring Boot, Flask, REST APIs, Microservices

Databases: MySQL, PostgreSQL, MongoDB, Vector DB (Pinecone)

Cloud/Tools: AWS (S3), Git, Docker

Frontend/UI: HTML, CSS, JavaScript (basic UI development)

WORK EXPERIENCE

Software Engineer (AI/ML)
Monocept Consulting Pvt. Ltd., Gurugram, India — *April 2024 – Present*

- Developed AI-powered solutions using Python and ML frameworks to automate business workflows.
- Built **Generative AI** and LLM-based agents using LangChain, LangGraph, and OpenAI for question answering and automation.
- Implemented **RAG pipelines** for knowledge retrieval and chatbot integration.
- Fine-tuned YOLOv8 model to detect Aadhaar info in documents with high accuracy.
- Developed backend services using Java and Spring Boot following microservices architecture.
- Led migration of databases from Oracle to PostgreSQL for Care Health Insurance ensuring **zero downtime**.

Associate Software Engineer
SuperSeva Global Services Pvt. Ltd., Bengaluru, India — *July 2022 – May 2023*

- Contributed to backend development using Java, Python, Spring Boot, and Flask.
- Implemented REST APIs and optimized code for service reliability.
- Worked with teams to automate internal workflows and enhance service delivery.

PROJECTS

Major Projects

- **PolicyAssist – Monocept:** Designed and developed an intelligent insurance assistant enabling users to **upload policy-related documents** and interact with them through natural language queries. Leveraged **Generative AI** and **RAG pipelines** with **LangChain** to deliver highly specific answers, improving customer experience and query resolution time.
- **Mono-DB-Agent – Monocept:** Engineered an **NLP-powered LLM agent** to translate natural language into MySQL queries using **LangChain** and **OpenAI**. Automated complex data retrieval workflows, reducing dependency on DBAs and increasing operational efficiency.
- **Mono-Aadhar-Masking – Monocept:** Developed a secure and automated pipeline for redacting sensitive Aadhaar numbers from uploaded documents using **OCR** and fine-tuned **YOLOv8** for precise detection, enhancing data security and compliance.

- **mPro – MaxLife Insurance:** Contributed to the backend development of a web application for insurance agents to capture user details and process policy purchases. Integrated new **Spring Boot APIs** and ensured secure storage with **AWS S3**, resolving performance bottlenecks and improving user experience.
- **DB Migration – Care Health Insurance:** Led a critical database migration from **Oracle** to **PostgreSQL**, refactoring existing **Java Spring** queries for compatibility, ensuring **zero downtime**, improved performance, and cost-efficiency.
- **PikmyKid – SuperSeva:** Handled and resolved customer queries via email and phone, acting as a primary point of contact. Ensured timely and accurate responses, improving customer satisfaction and maintaining quality support standards.

Data Science & AI/ML Development Projects

- **MoviesBot:** Developed a responsive chatbot using **HTML, CSS, JavaScript, Python (Flask)**, and **OpenAI**. Leveraged a **Retrieval Augmented Generation (RAG)** pipeline with **Pinecone** to answer movie-related questions efficiently.
- **SpringAIBot:** Created an AI-powered chatbot using **HTML, CSS, JavaScript, Java (Spring Boot)**, and **OpenAI**. This bot provides answers about Spring, Spring Boot, and Java topics and collects feedback.

Java-Oriented Projects

- **Electricity Bill Management System:** Built a Java-based application for electricity bill management enabling account creation, bill management, and payment tracking.
- **EasyRentHub:** Developed a housing system application using **Java, Hibernate**, and **MySQL**. This platform enables users to view and manage property rentals.
- **Trip Management System:** Designed a web application for online trip management using **Java, Spring Boot, MySQL, HTML, CSS**, and **JavaScript**.

General Web Development Projects

- **Purplle Clone:** Contributed to the development of a prominent online beauty and personal care platform using **HTML, CSS**, and **JavaScript**.

CERTIFICATIONS & COURSES

- Full Stack Development – View Certificate
- Complete Data Science Bootcamp – View Certificate
- LLM Engineering / Generative AI – View Certificate