

SUNNI KUMAR

+91 639454 40805 | contact.sunnyarya@gmail.com | [LinkedIn](#) | [GitHub](#) | Bengaluru, India

Summary

Software Engineer with 2+ years of experience building scalable backend systems, Generative AI applications, and distributed workflows. Experienced in Java, Spring Boot, Python, FastAPI, MySQL, MongoDB, AWS, and AI/LLM-powered systems including RAG pipelines and transformer-based NLP workflows. Strong foundation in Machine Learning, Deep Learning, distributed systems concepts, and production-grade software engineering through IIT Madras and real-world startup experience.

Education

Indian Institute of Technology Madras

B.S. in Data Science and Applications

May 2022 – Apr 2026

Chennai, India

- Relevant Coursework: Machine Learning, Deep Learning, Large Language Models, NLP, MLOps, Big Data, Database Systems, Software Engineering
- Hands-on exposure to NLP, Speech Processing, Computer Vision, and transformer-based architectures
- Diploma in Data Science (2024) and Diploma in Programming (2023)

Masai School

Full Stack Web Development - Backend Focused

Apr 2022 – Dec 2022

Bengaluru, India

- Completed intensive training in Java, Spring Boot, JavaScript, MySQL, and production-grade backend development
- Solved 1200+ DSA problems and built scalable full-stack applications

Work Experience

Sponte

AI Engineer (Founding Team)

Apr 2026 – Present

Bengaluru, India

- Architecting Generative AI workflows, multi-step LLM pipelines, and AI-powered enterprise solutions for production use cases.
- Building scalable FastAPI microservices integrated with vector databases including FAISS and pgvector for contextual reasoning over proprietary data.
- Working on prompt engineering, evaluation workflows, guardrails, and retrieval pipelines to improve model quality and reliability.
- Testing and validating product features before releases, improving platform stability, debugging workflows, and overall user experience.
- Collaborating with the founding team on AI product architecture, technical documentation, and deployment workflows.

Freshmenu

Software Development Engineer

Jan 2024 – Mar 2026

Bengaluru, India

- Built an AI-driven Financial Reconciliation Platform using **FastAPI**, **Next.js**, **SQL**, and **LLM workflows** to automate mismatch detection across Zomato, Swiggy, Magicpin, and Bitsila.
- Built an LLM-powered rules engine using **FAISS** and OSS GPT models to automate reconciliation workflows from partner agreement PDFs.
- Engineered a Procurement Automation System using **OpenCV** and AI-based invoice matching to reduce manual reconciliation workflows.
- Led ONDC integrations including **Magicpin** and **Bitsila**, implementing secure JWT authentication and asynchronous order processing workflows using **ActiveMQ**.
- Developed and maintained scalable backend services and REST APIs using **Java**, **Spring Boot**, **MySQL**, and **MongoDB** for production-grade systems.
- Optimized ordering APIs, resolved production timeout and null-pointer issues, and improved backend reliability through debugging and error handling improvements.
- Designed and developed backend services for a kitchen audit management system using **Java**, **Spring Boot**, and **MySQL**.
- Worked on HRMS full-stack modules, contributing to backend APIs, frontend workflows, and internal operational tooling.
- Improved internal reporting and admin portal workflows by optimizing backend services and operational visibility.
- Worked on scalable backend services and asynchronous integration workflows for production-grade distributed systems.
- Worked in Agile development workflows with collaborative debugging, CI/CD practices, feature releases, and production deployment support.

- Monitored cloud-based workflows on **AWS** and supported production debugging, deployment, and operational reliability improvements.

Freshmenu

Software Developer Intern

Sep 2023 – Dec 2023

Bengaluru, India

- Worked on backend engineering and API development tasks using **Java, Spring Boot, MySQL, and MongoDB**.
- Contributed to third-party integrations, feature enhancements, debugging, and production support workflows.
- Collaborated with engineering teams to understand business requirements and scalable backend system workflows.
- Gained hands-on experience in production-level software engineering, API integrations, and distributed backend systems.

Projects

AI Agent for Academic Guidance | *FastAPI, LangChain, pgvector, OpenAI*

- Built a full-stack Retrieval-Augmented Generation (RAG) system for context-aware academic guidance and Q&A workflows.
- Implemented JWT authentication, rate limiting, reranking, optimized chunking strategies, and vector-based retrieval pipelines.
- Deployed on AWS EC2 with Nginx, SSL, Docker, and GitHub Actions CI/CD workflows.

Spatio-Temporal Crime Prediction | *XGBoost, Scikit-learn, Pandas*

- Built a multi-class crime classification system on 800K+ records using temporal, spatial, and cyclical feature engineering.
- Applied GridSearchCV, k-fold validation, and model evaluation using precision, recall, and F1-score metrics.

Technical Skills

Languages: Java, Python, JavaScript, SQL, C++

Backend: Spring Boot, FastAPI, Node.js, Express.js, REST APIs, GraphQL

Frontend: React.js, Next.js, HTML, CSS, TailwindCSS

AI/ML: Machine Learning, Deep Learning, NLP, LLMs, Generative AI, RAG, LangChain, Hugging Face, PyTorch (Basics), XGBoost

Databases: MySQL, PostgreSQL, MongoDB, Redis, FAISS, pgvector

Cloud & Tools: AWS, Azure Fundamentals, Docker, Git, GitHub Actions, CI/CD, Linux, Nginx, Postman, Kibana

Core Concepts: Distributed Systems, System Design, API Integration, Microservices, Backend Development

Achievements

- Solved 1200+ DSA problems across coding platforms, strengthening algorithmic thinking and problem-solving skills.
- Selected among top students in the AlgoUniversity Accelerator Program and mentored by engineers from Google, Apple, and global product companies.
- Built production-grade AI and backend systems during undergraduate studies while working in fast-paced startup environments.