KOLLI GNANA JAYANTH REDDY

Sr Developer | Generative AI (Gen AI) Software Engineer | 4 Years Experience

Hyderabad, India | +91 7075062869 | jayanth.sde.fsd@gmail.com
LinkedIn: linkedin.com/in/jayanth-sde | GitHub: github.com/Jayanth-reflex | Portfolio: jayanth-sde.vercel.app

SKILLS

Languages: Python, TypeScript, JavaScript, Java

Generative AI and LLM: Large Language Models (LLMs), Retrieval Augmented Generation (RAG), Embeddings, Vector databases, Semantic search, Prompt engineering, Prompt templates, Prompt and model evaluation, Model quality assessment, Guardrails and Responsible AI, Agentic AI workflows (LangGraph)

Machine Learning and MLOps: Python, PyTorch, Hugging Face Transformers, Fine tuning and quantization (PEFT/LoRA, vLLM), Model training and inference, Weights and Biases (W&B), Model deployment, Monitoring, MLOps (Machine Learning Operations)

Data and Pipelines: Data pipelines, ETL (Extract, Transform, Load), Data ingestion, Apache Kafka (event driven pipelines), Structured and unstructured data, MySQL, MongoDB, Supabase

Backend and APIs: FastAPI, Spring Boot, Microservices architecture, Service oriented application design, RESTful and SOAP web services, API Gateway and API management, OAuth2 and JSON Web Token (JWT), Identity and Access Management (IAM), JPA/Hibernate

Security and Compliance: Role based access control (RBAC), Content Security Policy (CSP), HTTP Strict Transport Security (HSTS), Secrets management (AWS SSM), Data governance, Data privacy, Regulatory compliance, Secure handling of sensitive data

Frontend for AI Enabled Apps: React.js, Next.js, TypeScript, JavaScript, Data visualization and dashboards

Cloud, DevOps and Testing: Amazon Web Services (AWS), Microsoft Azure, Docker, Kubernetes, CI/CD (GitHub Actions, Jenkins), Git, Software configuration management (AWS SSM), JUnit5, Mockito, Integration testing, Splunk, New Relic, Sumo Logic

EXPERIENCE

iCIMS, Senior Developer — 09/2025 to Present

- Designed, developed, tested, and supported a production grade **Generative AI** platform on **AWS Bedrock** using **Retrieval Augmented Generation (RAG)**, **semantic and vector search**, and **Responsible AI guardrails** to support critical business processes, serving millions of users and cutting average application completion time by 60%.
- Improved model quality, safety, and accuracy through **prompt engineering**, **prompt templates**, and systematic **prompt and model evaluation**, reducing **LLM** hallucinations by 60% across production **agentic AI** workflows built on **LangGraph**.
- Fine tuned, evaluated, and deployed a domain specific **Small Language Model (SLM)** with **Python**, **PyTorch**, and **Hugging Face Transformers**, supported by end to end **MLOps** on **Docker** and **Kubernetes** with **Weights and Biases (W&B)** observability, reducing third party LLM costs by 80%.
- Secured the platform with **role based access control**, unifying modern **Identity and Access Management (IAM)** and legacy **JSON Web Token (JWT)** users into a single, consistent authorization model.
- Hardened application security by enforcing **Content Security Policy (CSP)** and **HTTP Strict Transport Security (HSTS)** and eliminating hardcoded secrets through centralized **AWS Systems Manager (SSM) Parameter Store** management.
- Collaborated **cross functionally** with engineering, data science, and product teams, authoring **reusable implementation patterns** and technical documentation, mentoring 3 developers, upskilling 15+ engineers on **RAG** and **agentic AI**, and leading proofs of concept (**Claude**, **Model Context Protocol (MCP) servers**) that shaped the company's AI strategy.

Wipro, Software Engineer — 09/2022 to 03/2025

Hyderabad, India

- Developed a **Retrieval Augmented Generation (RAG)** chatbot grounding **Large Language Model (LLM)** responses in enterprise content (company, job, and job description data), enabling accurate self service information retrieval on demand.
- Built an **enterprise systems integration** layer across 10+ downstream systems using **REST and SOAP web services** and **Apache Kafka** pipelines, handling **ETL (Extract, Transform, Load)** and ingestion of structured and

unstructured data.

- Designed and developed 4 **Java Spring Boot microservices** supporting 50,000+ concurrent users, containerized and orchestrated with **Docker** and **Kubernetes**, applying **Software Development Life Cycle (SDLC)**, **solution architecture**, system design, and **service oriented application design**.
- Implemented **API management** with **Azure API Management** secured by **OAuth2 and JWT**, reducing security incidents by 60%, and engineered **React.js** and **TypeScript** dashboards integrating 25+ APIs for **data visualization**, increasing operational efficiency by 45%.
- Optimized **MySQL** schemas with **JPA/Hibernate**, reducing response times by 35%, built **JUnit5** and **Spring Test** suites to 90%+ code coverage, and led end to end **monitoring** and **IT operations** with **Splunk** and **New Relic (APM)**.

Wipro, Java Full Stack Developer Trainee **02/2022 to 08/2022**

- Built enterprise **microservices** and **React.js** and **TypeScript** components, and implemented **REST and SOAP services** with automated testing (**JUnit** and **Mockito**).

CERTIFICATIONS AND ACHIEVEMENTS

- **Certifications:** Google Generative AI Certified, AWS Certified, Microsoft Azure Certified
- **Vista Hackathon Winner:** Clinched a Vista backed hackathon with an agentic AI system that autonomously shortlists, interviews, and recommends candidates across 1M+ applicant records, slashing manual research and screening effort for recruiters and candidates by 90%.
- **Wipro "Rising Star" Award:** Recognized for standout technical leadership and high impact AI integration delivery.

PROJECTS

- **Domain Specialized SLM Fine Tuning and Evaluation** *(Python, FastAPI, Hugging Face Transformers, PyTorch, PEFT/LoRA, Weights and Biases)*: Fine tuned and evaluated a domain specific model deployed behind a Llama Guard safety gateway with guardrails to keep responses safe and on policy for regulated knowledge domains.
- **RAG Recommendation Platform on Gemini** *(Java, Spring Boot, RAG, Google Cloud)*: Built a scalable, secure Retrieval Augmented Generation (RAG) system with vector and semantic search for personalized recommendations.
- **Healthcare Domain Disease Tracking Platform** *(Java, Spring Boot, microservices, MySQL, MongoDB, React.js, TypeScript, Docker)*: Built a healthcare monitoring platform with a React frontend and microservices backend supporting 20,000+ users.

EDUCATION

Bachelor of Technology, JNTUH (MRCET), Secunderabad **07/2018 to 06/2022**