

PROFESSIONAL SUMMARY

AI application developer with **2+ years** across Oracle Fusion SCM/SCP, agentic AI workflows, RAG systems, and cloud-native deployments. Built **multi-agent** retrieval and automation workflows using **Python, LangChain, LangGraph, vector databases, FastAPI, Kubernetes, OCI/OKE, KServe, and Milvus**. Google Summer of Code 2026 maintainer for Kubeflow and winner of Oracle Gen AI Hackathon among **33 teams / 300+ employees**.

TECHNICAL SKILLS

Languages: Python, Java, C++, JavaScript, TypeScript, SQL
AI / ML: LangChain, LangGraph, RAG, MCP, Kagent, agent evals, Ollama, TEI embeddings, FAISS, ChromaDB, Milvus, Flowise, PandasAI
Backend & APIs: FastAPI, Node.js, Express.js, REST APIs, Java ADF, Postman, Prisma
Frontend: React, Next.js, Tailwind CSS, ShadCN, ReactFlow
Cloud / DevOps: Docker, Kubernetes, OCI, OKE, KServe, Kubeflow Pipelines, GitHub Actions, CI/CD
Databases: PostgreSQL, MySQL, Elasticsearch, Supabase, Firebase

EXPERIENCE

AI Application Developer

Jun 2024 - Present

Bengaluru, India

- Built context-rich AI integrations for **Oracle Fusion SCM Cloud** across **3 customer-facing workflows**—catalog lookup, order review, and planner assistance—using **Python, LangChain, LangGraph, RAG, embeddings, and vector databases**.
- Developed an **MPN and part-matching pipeline** combining fuzzy matching, semantic retrieval, clustering, and web-search fallback over manufacturer and retailer catalogs, improving match coverage for asset-mapping workflows by **~30%**.
- Implemented a **release/review order-classification agent** with **ESS batch processing**, rule-evaluation guardrails, exception routing, and human-review handoff, enabling bulk processing of **10K+ order lines** per run.
- Created RAG-based planner-assist flows over **SCP Exceptions, Notes, and pivot/tabular data**, reducing repeated manual investigation by **~40%** through source-grounded answers, generated notes, and reusable retrieval filters.

Open Source Maintainer - Kubeflow

May 2026 - Present

Remote

- Selected for GSoC 2026 to expand kubeflow/docs-agent into an **Agentic RAG reference architecture** for Kubeflow users, improving access to docs, troubleshooting context, and implementation examples.
- Built **multi-index retrieval** across GitHub issues, application code, Kubernetes manifests, and **1K+ Markdown docs**, producing **10K+ searchable chunks** with path, product-area, version, and source metadata.
- Integrated issue history, docs, and source-code context into agent responses, cutting multi-page manual lookup for Kubeflow troubleshooting from **minutes to seconds**.
- Developed Kubeflow Pipelines ingestion workflows, **50+ CI-backed retrieval tests**, and **KServe deployment** patterns on OCI/OKE for repeatable cloud-native agent serving.

Application Developer Intern

May 2023 - Jul 2023

Hyderabad, India

- Developed **Copy Data** functionality for lead-time columns in Oracle SCM during a 3-month internship.
- Optimized retrieval and update workflows using **Elasticsearch** queries and **PATCH-based API** updates.
- Added **Linear Regression** trend analysis to Oracle JET charts for supply-chain visualization and forecasting use cases.

PROJECTS

MacBatch - Distributed Batch Inference on Idle Apple Silicon

2026

Product Source

- Built an open-source **batch inference platform** that distributes latency-tolerant AI jobs—**embeddings, OCR, and classification**—across a pool of idle Apple Silicon machines running open models locally through **Ollama**.
- Designed a **FastAPI** control plane modeling jobs → shards → tasks, with **600s worker leases**, automatic reclaim and reassignment on failure, and pull-based workers that join behind NAT over outbound HTTPS only.
- Benchmarked **252,686 embedding items/hour** on a single node and a **2.8×** throughput gain from shard batching alone, publishing a **cost model** against GPU rental with reproducible methodology.

CrackRound - Agentic AI Mock-Interview Platform

2025

Website

- Built an end-to-end **agentic AI interviewer** from scratch to production with **5 streaming personas** across system-design, conceptual, and behavioral rounds, plus **LLM-vs-LLM simulation** that builds a candidate persona from a resume and self-plays full interviews.
- Engineered a real-time voice loop with streaming STT → GPT-4o → streamed TTS over WebSocket at **~1.5s response latency**.
- Integrated a live DSA code judge and system-design whiteboard into the LLM context, then productionized scoring with JSON-schema evaluation, **5-dimension feedback**, auth, credits, and **\$2/session** cost guardrails.

EDUCATION

Bachelor of Technology in Mechanical Engineering

2020 - 2024

National Institute of Technology, Warangal

CGPA: 7.34

ACHIEVEMENTS

- Accepted for **Google Summer of Code 2026** under Kubeflow for an Agentic RAG project; speaker at **Kubeflow Community Showcase 2026**.
- Won **Oracle Gen AI Hackathon: 1st among 33 teams, top 5 among 300+ employees**; Runner-up, **Tri-NIT Hackathon 2024**, Backend Developer track.
- Solved **600+ coding problems** across LeetCode, GeeksforGeeks, CodeChef, and Codeforces; conducted **15+ competitive programming contests**.