

Chetan Anand

+91 88280 73690 | chetan.anand710@gmail.com | linkedin.com/in/chetanand

SUMMARY

Software Engineer with 5+ years of experience building **scalable AI, backend, and cloud-native systems**. Experienced in **RAG, LLM integration, high-concurrency APIs, CI/CD, observability**, authentication and latency optimization across Cloud infrastructure.

WORK EXPERIENCE

Software Engineer (IC2)

Oracle

07/2024 – Present

Bangalore, Karnataka

- Built a Retrieval-Augmented Generation (RAG) application with **LlamaIndex** for PDF/CSV ingestion and vector indexing, enabling context-grounded responses with source attribution.
- Integrated OCI **Generative AI** services with **LangChain** and pretrained **Cohere** models for embeddings, semantic search, text classification, and contextual response generation.
- Engineered a high-concurrency **FastAPI** backend for RAG response generation, supporting **1,000 concurrent users** and reducing load-test request failures from **40% to 0.6%**.
- Built a **hybrid CI/CD deployment pipeline** supporting backend-only rsync updates and full **Docker**/Podman image deployments through **Jenkins**, enabling near-zero-downtime upgrades and removing manual release steps.
- Implemented secure application authentication using **Oracle IDCS** and **OAuth 2.0**, including client-credentials flows, JWT-based identity/context extraction, token management, and protected API access.
- Implemented **OpenTelemetry** distributed tracing with OCI APM, instrumenting **HTTP requests, database operations, retrieval**, and **GenAI** calls to provide end-to-end latency breakdowns and failure diagnostics.
- Designed enterprise provisioning flows spawning WTSS pods to onboard **4000** enterprise tenants into the application.
- Implemented **Dedicated AI Cluster** - based optimizations to reduce application latency by **6 sec** and improve responsiveness in latency-sensitive request flows.

Software Developer

IBM

08/2021 – 07/2024

Bangalore, Karnataka

- Built an N-gram (bigram) language model and text-classification pipeline, improving multiclass classification accuracy by 14%.
- Applied SMOTE to imbalanced datasets and trained models with Watson NLU to reduce class bias and improve prediction quality.
- Orchestrated CI/CD pipelines for AI applications using IBM Cloud, Git, and Docker.
- Prompt-tuned LLMs for sentiment analysis to capture and classify user feedback.
- Optimized pagination and page-loading behavior, reducing page load time from 58 seconds to 4 seconds.
- Improved conversational-flow predictions using intent and entity modeling; automated login and password-update workflows with Selenium and Python.

SKILLS

Skills: Python, FastAPI, Django, Flask, Generative AI, RAG, LlamaIndex, LangChain, NLP, Prompt Engineering, OpenTelemetry, OCI Generative AI, Oracle IDCS, OAuth 2.0, Jenkins, CI/CD, Docker, Podman, Git, Jira, IBM Cloud, Selenium

EDUCATION

Bachelor of Technology (B.Tech) in Computer Science

Kalinga Institute of Industrial Technology (KIIT), Bhubaneswar

2017 – 2021

8.4 CGPA

CERTIFICATIONS

Oracle Cloud Infrastructure 2024 Generative AI Certified Professional