Yashovardhan Tiwari

+91-7073215996 | yt282003@gmail.com | linkedin.com/in/yashovardhan-tiwari | github.com/codezeewrangler

Education

Vellore Institute of Technology Bhopal, MP
Bachelor of Technology in Computer Science and Engineering; CGPA: 8.08/10.0 *Expected May 2026*

Experience

AI & Full-Stack Engineer Jan 2025 – Present
Self-Directed AI Systems Engineering & Open Source *Remote*

- Independently designed and shipped **2 production-grade** AI applications using LangGraph, LangChain, and FastAPI – handling full architecture, deployment, and iteration without team oversight.
- Built RAG pipelines with FAISS and ChromaDB; semantic similarity retrieval returned contextually relevant results where exact keyword search failed on **3 out of 5** test query categories during self-evaluation.
- Containerized and deployed backend services on Render using Docker; measured average API response times of **280–320ms** under personal load testing with **up to 20 concurrent requests**.
- Added Redis caching for repeated LLM calls, cutting redundant API round-trips on identical prompts – observed **2x** response speed improvement on cache hits during local benchmarking.

Projects

InsightATS Resume Builder | *Python, FastAPI, LangChain, React, Docker, FAISS* Live Demo

- Engineered a full-stack ATS resume analysis platform that parses resumes and job descriptions through a LangChain LLM pipeline, computing structured keyword alignment scores between candidate profiles and job requirements.
- Built a multi-agent workflow with dedicated agents for skill extraction, gap analysis, and suggestion generation – transforming unstructured resume text into actionable improvement reports.
- Implemented semantic similarity scoring using Hugging Face sentence embeddings and FAISS vector search, enabling context-aware skill matching beyond exact keyword lookup.
- Deployed FastAPI backend with Docker on Render, integrating Redis response caching to achieve consistent **sub-300ms** API latency under standard load.
- Secured the platform with JWT-based authentication, CORS policies, and per-endpoint rate limiting to meet production reliability and data privacy standards.

Nexus AI Agent | *Python, LangChain, FastAPI, Redis, Hugging Face* Live Demo

- Integrated LangChain tool calling and persistent contextual memory, enabling the agent to autonomously break complex user queries into sequential, dependency-aware sub-tasks.
- Optimized inference throughput by implementing Redis-backed response caching and prompt batching, significantly reducing redundant LLM calls on repeated or similar queries.
- Engineered an async task queue with automatic retry logic and exponential backoff, improving resilience against API timeouts and transient model provider failures.
- Secured all endpoints with JWT authentication, rate limiting, and CORS protection; implemented structured JSON logging for end-to-end request traceability in production.

Technical Skills

Languages & Scripting: Python, Java, SQL, Bash
Backend Frameworks: FastAPI, LangChain, LangGraph, Node.js
AI / LLM: RAG Pipelines, LLM Orchestration, Prompt Engineering, Hugging Face Transformers
Vector Search & Storage: FAISS, ChromaDB, Embedding Pipelines
Databases & Caching: Redis, PostgreSQL, MongoDB
DevOps & Cloud: Docker, AWS, Render, CI/CD

Certifications

OCI AI Foundations – Oracle (Sept 2025) | Cloud AI Infrastructure, Model Deployment
Generative AI Professional – Oracle (Oct 2025) | LLM Integration, Prompt Engineering
Machine Learning Specialization – Coursera, Andrew Ng / Stanford (Dec 2023)