Aayush Yadav

AI Engineer — Generative AI Engineer

7021907344 — aayushdyadav2003@gmail.com
github.com/03ADY — aayush-yadav-portfolio.vercel.app

Summary

AI Engineer with hands-on experience building and deploying **production-grade Generative AI systems**, including **Retrieval-Augmented Generation (RAG)** and multimodal inference pipelines. Strong background in backend engineering and scalable AI architectures, with a focus on **reliability, security, and maintainability** in enterprise environments.

Professional Experience

AI Engineer — Jobihood Technologies (Remote) *Nov 2025 – Present*

- Built an AI-powered resume intelligence system using Gemini LLMs to extract structured candidate profiles from unstructured PDF resumes, including OCR fallback.
- Architected a multi-tenant vector search system with isolated databases per client, ensuring strict data separation and enterprise-grade security.
- Implemented a hybrid retrieval pipeline combining semantic vector search with LLM-based intent parsing for accurate filtering by skills, experience, and location.
- Developed logic to distinguish professional experience from academic projects, enabling accurate per-skill seniority estimation.

Python Backend Developer Intern — NDT Reflection Engineers (Remote) *Oct 2024 – Mar 2025*

- Developed scalable REST APIs using FastAPI and Flask, supporting authentication and core business workflows.
- Managed PostgreSQL databases via SQLAlchemy ORM, optimizing query performance and data integrity.
- Containerized backend services using Docker, improving deployment reliability and CI/CD workflows.

Selected Projects

AI Assistant (Aura) — Enterprise RAG Chatbot

- Built a production-grade RAG system using FastAPI, LangChain, Gemini 2.5 Flash, and ChromaDB for accurate, context-aware knowledge retrieval.
- Designed secure ingestion pipelines for PDF, DOCX, CSV, and TXT documents with automated vector indexing.
- Reduced information retrieval time and improved decision-making through grounded AI responses.

Resume Intelligence & AI-Powered Recruitment Engine

- Developed an end-to-end GenAI pipeline for structured resume parsing using LLMs and OCR fallback.
- Implemented semantic and intent-based search to enable precise candidate filtering.
- Designed with enterprise scalability and data security as primary constraints.

Multi-Modal Inference Engine (Vision & Speech)

- Engineered a real-time system integrating speech-to-text (Whisper) and object detection (YOLOv5).
- Optimized GPU inference using TensorRT and CUDA, achieving up to 50% faster processing.
- Built asynchronous APIs with WebSocket support for real-time streaming and reduced latency.

Skills

Generative AI & NLP: RAG, Prompt Engineering, Embeddings, Vector Search, Gemini LLMs, LangChain
Backend & Systems: Python, FastAPI, Flask, REST APIs, WebSockets, Docker, CI/CD
Machine Learning: Deep Learning, Computer Vision, Time-Series Modeling, Reinforcement Learning, SHAP, LIME
Data & Infrastructure: PostgreSQL, SQLite, ChromaDB, GPU Inference, TensorRT, CUDA
Tools: PyTorch, TensorFlow, Whisper, YOLOv5, NumPy, Pandas, Git

Education

B.Tech — Artificial Intelligence & Data Science (Honors in Cybersecurity)
A.C. Patil College of Engineering, Navi Mumbai *2021 – 2025*
GPA: 8.4