Jasveen Singh Sahani

Toronto, ON jasveen1800@gmail.com 4379860030 LinkedIn GitHub Portfolio

SUMMARY

Junior AI Engineer focused on building production-ready LLM and agentic AI systems. Experienced in designing end-to-end pipelines using LangGraph, AWS Bedrock, and OpenAI, including RAG systems, semantic search, and scalable FastAPI backends. Strong background in deploying containerized AI applications on AWS (ECS Fargate) with CI/CD automation.

SKILLS

Languages: Python, Java, SQL
AI / Machine Learning: Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), semantic search, prompt engineering, hallucination mitigation, scikit-learn, XGBoost, LightGBM, pandas, NumPy
AI Frameworks & Tools: LangGraph, LangChain, LlamaIndex, ChromaDB, FAISS
API & Backend: FastAPI, REST APIs, backend service development, Pydantic validation, async API handling
Cloud & DevOps: AWS Bedrock, AWS ECS Fargate, AWS ECR, Docker, Docker Compose, GitHub Actions (CI/CD), Git

PROJECT EXPERIENCE

InsightStream – Multi-Agent AI News Digest System GitHub
Feb 2026 – Apr 2026

- Built and deployed a production-grade multi-agent (agentic AI) pipeline using LangGraph with 6 specialized agents orchestrating Claude Sonnet 4.5, Llama 3.3 70B (AWS Bedrock), and GPT-5.1/5.2 (OpenAI).
- Engineered a scalable news ingestion system monitoring 41 RSS feeds, processing 400+ articles per run with semantic deduplication (cosine similarity tuning) and diversity-aware ranking across 8+ user-defined categories.
- Built and delivered personalized AI-curated daily news digests based on natural language preferences, eliminating manual filtering across 40+ sources.
- Deployed on AWS ECS Fargate using Docker with CI/CD via GitHub Actions, enabling automated builds, ECR pushes, and production deployments with 5-minute end-to-end runtime.

Neural Semantic Job Search Engine GitHub
Jan 2026 – Feb 2026

- Built an LLM-powered AI application that matches resumes with Canadian job postings using Meta Llama 3.3 via AWS Bedrock with a FastAPI backend and Streamlit interface.
- Implemented resume parsing and data preprocessing pipelines to extract structured candidate information from PDF resumes for model-based job relevance scoring.
- Designed a hybrid ranking pipeline combining rule-based filtering with LLM-driven semantic analysis, improving job matching accuracy compared to traditional keyword approaches.
- Deployed the system using Docker and Docker Compose, enabling containerized AI inference services and real-time job ingestion via the Adzuna API.

FinTech RAG Copilot – Regulatory Compliance AI Assistant GitHub
Jan 2026

- Developed a Retrieval-Augmented Generation (RAG) system using LangChain and AWS Bedrock (Claude Sonnet) to answer financial regulatory questions from OSFI and cybersecurity documents.
- Built a document ingestion and preprocessing pipeline using PyPDF and LangChain to generate Amazon Titan embeddings stored in ChromaDB for semantic retrieval.
- Implemented a FastAPI REST API to orchestrate retrieval and LLM generation, enabling real-time question answering with citation-backed responses.
- Containerized backend services and UI using Docker and Docker Compose, supporting scalable deployment of AI-powered compliance tools.

EDUCATION

Bachelor of Science in Computer Science
York University 2026

CERTIFICATIONS

Generative AI with Large Language Models DeepLearning.AI
ChatGPT Prompt Engineering for Developers DeepLearning.AI
AI & ML Bootcamp Udemy