DHIAN SINGH

Generative AI Engineer

Location: Mohali, Punjab, India
Email: dhiansingh906907@gmail.com
Phone: +91 99143 35981
LinkedIn: https://linkedin.com/in/dhian-singh-15019b230
GitHub: https://github.com/dhian-singh

PROFESSIONAL SUMMARY

Results-driven Generative AI Engineer with 6 years of progressive experience architecting, developing, and deploying enterprise-scale artificial intelligence solutions. Distinguished expertise in Retrieval-Augmented Generation (RAG) systems, agentic AI frameworks, conversational AI, AI chat systems, computer vision, and speech recognition. Demonstrated proficiency in Python development, large language model optimization, and multi-cloud deployments across AWS, GCP, and Azure. Skilled in leveraging CUDA programming with NVIDIA GPUs, PyTorch, and Torchvision for accelerated deep learning workflows, as well as optimizing vector database embeddings with GPU acceleration for high-speed similarity search and retrieval. Proven track record of leading cross-functional teams, streamlining MLOps pipelines, and delivering transformative AI solutions that drive measurable business impact across diverse industry verticals.

TECHNICAL EXPERTISE

Programming & Development: Python, R, Git/GitHub, FastAPI, Dango,React,Next js,RESTful API Design, Microservices Architecture

AI & ML Domains: Data Science,Deep Learning , NLP, LLMs, RAG Implementation,Generative AI, Agentic AI (CrewAI, AutoGen, AGNO,Langpaph), Conversational AI, AI Chat Systems, Computer Vision (OpenCV, YOLO, image classification & detection, OCR), Speech Recognition (ASR, TTS), Deep Learning, LangChain, LangGraph, Model Context Protocol (MCP)

Libraries & Frameworks: TensorFlow, PyTorch, Torchvision,OCR , HuggingFace, OpenCV, YOLO, NumPy, Pandas, Scikit-learn, FAISS, ChromaDB,Mongo DB.FIAS,PIcorn.

GPU & Acceleration: NVIDIA CUDA RTX GPUs, CUDA Programming, GPU-accelerated Vector Embeddings & Similarity Search

Cloud Platforms: AWS (SageMaker, Bedrock, Lambda, S3), GCP, Azure

Core Competencies: Technical Leadership, Cross-functional Collaboration, Problem-Solving, AI System Architecture, MLOps Automation

PROFESSIONAL EXPERIENCE

AI ML Developer

Zoptal Solutions Pvt. Ltd., Mohali, India (June 2025 – 31 July 2025)

- Spearhead the development of cutting-edge generative AI solutions integrating computer vision, NLP, and speech recognition.
- Collaborate with cross-functional teams to define AI product roadmaps, technical architectures, and deployment strategies.

Senior Generative AI Engineer

Techlive Solutions, Mohali, India (Nov 2024 - Apr 2025)

- Architected and deployed AI systems for production-grade conversational AI platforms.
- Conducted training sessions for engineering teams on LLM fine-tuning, latency optimization, and RAG pipelines.

Generative AI Trainer

UCT & IIT Roorkee (Aug 2024 - Sep 2024)

- Designed and delivered workshops for 200+ participants on RAG systems, agentic AI, and cloud deployment.

- Provided hands-on training in building AI chatbots with LangChain and LLM APIs.

Python Generative AI Engineer

Techedo Technology (Aug 2023 - Mar 2024)

- Developed and deployed 15+ LLM-powered applications, achieving 30% reduction in API costs through AWS optimization.
- Enhanced system throughput by reducing latency by 25% in chatbot and RAG implementations.

Python AI Developer

CBIITTS Technology (Apr 2018 - Apr 2023)

- Delivered 25+ AI models across NLP, CV, and speech-to-text applications.
- Implemented MLOps pipelines enabling faster model deployment cycles.
- Mentored 8 junior developers on AI/ML best practices.

SIGNATURE PROJECTS

Enterprise Astrology RAG System

Tech Stack: RAG, AGNO, Groq, Streamlit, FastAPI, PostgreSQL

- Built a multimodal AI platform combining text and date-of-birth inputs for personalized astrology predictions.
- Integrated persistent memory and custom embeddings for highly relevant, context-aware answers.

RAG Chatbot & Knowledge Assistant

Tech Stack: GPT, Ollama, LangChain, AWS Lambda, Wikipedia API, Google Search API

- Developed a real-time knowledge assistant capable of pulling and summarizing live data from the web.
- Implemented context chaining to maintain coherent multi-turn conversations.

AI-Powered Document Intelligence Platform

Tech Stack: LangChain, FAISS, ChromaDB, LLMs, AWS Lambda, OCR

- Created a serverless document analysis system with semantic search and document summarization.
- Integrated OCR and speech-to-text for processing scanned PDFs and audio-transcribed documents.

EDUCATION

Bachelor of Technology in Computer Science & Engineering

Punjab Technical University, Jalandhar (2013 - 2016)

ADDITIONAL QUALIFICATIONS

Languages: English, Hindi, Punjabi

Professional Development: Ongoing learning in AI, MLOps, leadership

Personal Interests: Technical literature, culinary arts, dance, motorcycle touring