Krishan Kumar

+91 6280457257 — krishana85289@gmail.com — linkedin.com/in/krishan-kumar — github.com/krishana85289

SUMMARY

AI Engineer with 4+ years of experience in AI/ML/Generative AI engineering, specializing in **Large Language Models (LLMs)** and enterprise-scale AI solutions. Experienced in designing and deploying **RAG pipelines, agentic chatbots, data extraction agents, and AI-powered decision agents**. Skilled in developing and optimizing **high-impact prompts, agentic workflows, and LLM API integrations** (OpenAI, vLLM, open-source models) to build scalable enterprise services. Proficient in **Python, cloud platforms (AWS preferred, GCP)**, containerization (Docker, Kubernetes), and CI/CD for production-grade AI deployments.

PROFESSIONAL EXPERIENCE

22nd Century Software Solutions
Apr 2025 – Present

AI Engineer — *Mohali, Punjab*

- Architected and deployed **LLM-powered agentic workflows** using LangChain and LangGraph to automate complex document understanding tasks, including **IRS document extraction, classification, and analysis**.
- Led the development of an AI-powered proposal generation platform with **Generative AI agents**, reducing manual RFP preparation effort by **70%** and improving proposal success rates.
- Developed a machine learning model to predict issue resolution time from historical support data, improving forecasting accuracy and optimizing resource allocation.

TMotions Global Limited
Jan 2021 – Apr 2025

AI Engineer — *Mohali, Punjab*

- Pioneered the adoption of **AI/ML initiatives** within the company, moving from proof-of-concepts (POCs) to full-scale production systems, which established AI as a key business growth driver.
- Developed an automated service to extract structured data from unstructured PDFs using fine-tuned open-source models, transforming raw documents into XML format. Eliminated manual processing and achieved a **99% cost reduction**.
- Scaled open-source LLMs, vector databases, and embedding models on AWS EC2 to create **15+ secure, client-specific chatbots**. Delivered a cost-effective solution deployed on client-owned servers, ensuring data security and compliance while avoiding recurring SaaS fees.
- Designed and deployed a recommendation engine for a medical company to match pharmaceutical firms with suitable doctors for research studies, cutting outreach costs by **50%** and improving recruitment efficiency.
- Developed a deep learning–powered AI search engine, improving search accuracy and boosting customer satisfaction by **35%**.
- Built an AI agent using **Vision-Language Models (VLMs)** and Retrieval-Augmented Generation (RAG) to review traffic challans with photo/video evidence. Automated document and proof analysis reduced manual workload by **80%** and lowered operational costs.

TECHNICAL SKILLS

Programming: Python (Pandas, NumPy, Scikit-learn, PyTorch, TensorFlow, Keras), FastAPI, Flask

Generative AI / LLMs: RAG Pipelines, Prompt Engineering, LangChain, LangGraph, OpenAI APIs, Gemini APIs, Hugging Face Transformers, Fine-tuning, Vision-Language Models (VLMs), Open-source LLMs (LLaMA, Gemma, Mistral, etc.)

Machine Learning & NLP: Classical ML, Deep Learning, Computer Vision, OCR, NLP

Cloud & MLOps: AWS (SageMaker, Lambda, EC2), GCP (Vertex AI), Azure ML, Docker, Kubernetes, CI/CD

Tools: Git, Linux, ChromaDB, Pinecone, vLLM, Hugging Face Hub

PROJECTS

AI-powered Decision System | ***GenAI, LangGraph, Python*** May 2025

- Built an AI-powered decision system that evaluates supporting evidence (photos, videos, documents) for traffic violations and determines whether a challan should be waived or upheld.
- Automated a traditionally manual review process using GenAI agents, reducing the need for 20–30 staff members from 50 and significantly cutting operational costs.
- Integrated LangGraph to design intelligent workflows that assess proof authenticity, rule violations, and context — simulating human judgment in traffic violation disputes.
- Enabled near real-time resolution of traffic challans, improving transparency, accuracy, and efficiency in law enforcement workflows.

Medical Data Chatbots for Pharmaceutical Studies | ***Python, OpenAI LLM, ChromaDB*** Apr 2025

- Created an intelligent chatbot designed to streamline data retrieval and interactions with medical datasets, boosting efficiency for pharmaceutical research and study.

Automated PDF Data Extraction Service | ***Python, Deep Learning, Open-source LLM*** Jan 2025

- Developed an automated service to extract structured data from unstructured PDFs using fine-tuned open-source models, transforming raw documents into XML format.
- Reduced manual data extraction workload by 100%, achieving a 99% cost reduction by fully automating the process and minimizing human intervention.
- Enhanced efficiency and accuracy by streamlining the data extraction pipeline, reducing the need for manual data handling and significantly lowering operational costs.

EDUCATION

Chandigarh Engineering College Sep 2016 – May 2020
Bachelor of Technology (B.Tech) in Electronics and Communication Engineering *Mohali, Punjab*

CERTIFICATIONS

Microsoft Certified: Azure Data Scientist Associate 2024

Intellipath Master's Program in Data Science and AI 2024 – 2025 (Ongoing)