# JAGDISH

## LLM AND DEEP LEARNING ENGINEER

Rewari, HR 123401 • 89026474629 • jagdishlamba72@gmail.com • **LinkedIn:** linkedin.com/in/jagdishlamba • **WWW:** https://jagdishlamba.github.io

## Professional summary

Results-driven Deep Learning and LLM Engineer with hands-on experience in designing, developing, and deploying end-to-end AI solutions across computer vision, natural language processing, and edge AI. Proven expertise in building scalable ML pipelines, training models using multi-GPU setups, and deploying optimized deep learning models on resource-constrained devices like NVIDIA Jetson and Raspberry Pi. Strong background in Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), and real-time video analytics using GStreamer and DeepStream. Adept at working across the full ML lifecycle—from data engineering and model development to evaluation, optimization, and production deployment. Known for a strong ownership mindset and the ability to deliver high-performance AI systems in real-world environments.

## Skills

| | |
|---|---|
| Multi-GPU Training | Transfer learning |
| ONNX, TensorRT, model quantization | MLOps, experiment tracking |
| LLM fine-tuning | Prompt engineering |
| Retrieval-Augmented Generation (RAG) | semantic search |
| LangChain, Hugging Face Transformers, vector databases | Real-time video analytics |
| OpenCV, YOLO | Model deployment on edge devices |
| DeepStream and Gstreamer | MLFlow and Airflow |
| Django, Flask and fastAPI | Anaconda |

## CERTIFICATION

Python Zero to Hero Bootcamp: Udemy

Python for Data Science: CognitiveAI (IBM)

Plotly and Dash in Python: Udemy

Django Bible Course: Udemy

Pre-Program Preparatory Content course for AI & ML

## Work history

**LLM and Deep Learning Engineer**, 07/2019 to Current
**Central Government** – Indore

Worked as a core member of the AI/ML team, leading the full lifecycle of machine learning and deep learning projects—from problem definition, dataset creation, to model deployment and monitoring in production. Specialized in both large language models (LLMs) and computer vision systems, with hands-on experience implementing custom and open-source architectures.

- Designed and built scalable ML pipelines using PyTorch and TensorFlow, leveraging multi-GPU training strategies to accelerate model convergence on large datasets reduced training time by 60%.
- Implemented Retrieval-Augmented Generation (RAG) pipelines using vector databases and LLM frameworks such as LangChain and Hugging Face Transformers to build context-aware, domain-specific language applications reduced the manual effort by 50%.
- Deployed lightweight and optimized deep learning models on edge devices, including NVIDIA Jetson Nano and Raspberry Pi, for real-time inference tasks. Worked with TensorRT and ONNX to reduce latency and model size.
- Developed real-time video analytics solutions using GStreamer and NVIDIA DeepStream SDK, integrating multiple AI models for multi-stream object detection and tracking applications (25 cameras in one instance).
- Responsible for the full MLOps lifecycle, including model versioning, performance tracking, experiment management etc.
- Developed front end UI for model deployment and testing using Gradio, Django, NextJS and flask

## Education

**Master of Technology**: Data Science & AI, 09/2027
**IIIT** – Dharwad

**Master of Science**: AI & ML, 11/2024
LJMU - UK

**PG Diploma**: AI & ML, 08/2023
**IIIT** - Bengaluru

## PROJECTS

**1. Multi-Feed Real-Time Object Tracking System**

Developed a real-time object tracking solution using OpenCV (compiled with CUDA) and GStreamer for high-performance multi-camera feeds. Designed custom object classes to detect intrusions and trigger real-time alerts. Optimized for low-latency performance on GPU-enabled systems.

**2. Offline Real-Time Automatic Number Plate Recognition (ANPR)**
Built a robust ANPR system trained on a custom vehicle dataset, capable of recognizing number plates in real-time without internet connectivity. Achieved high accuracy in various lighting and angle conditions using custom OCR and detection pipelines.

**3. Real-Time Multi-Feed Face Recognition & Logging System**
Implemented a face recognition system for tracking individuals across multiple camera feeds. Generated structured logs for entry and exit events, supporting offline operation and high accuracy across diverse lighting conditions.

**4. Offline LLM-Based Document Summarization**
Designed an offline pipeline for summarizing long-form documents using fine-tuned LLMs. Built a lightweight, local-first architecture suitable for environments with data privacy constraints or limited internet access.

**5. Domain-Specific Offline Question Answering Chatbot**
Developed a chatbot using Retrieval-Augmented Generation (RAG) and semantic search on domain-specific datasets. Operated entirely offline, delivering accurate, context-aware responses with fast inference times.

**6. Edge Device Object Detection & Tracking**
Deployed real-time object detection and tracking models on NVIDIA Jetson and Raspberry Pi. Optimized models using TensorRT and ONNX for efficient inference on low-power edge devices.

**7. Natural language to SQL generation**
Deployed a chatbot for conversation in plain English language to get real time results from RDBMS in place of complex SQL queries making it very practical and enhance UX.

## Git Profile

- https://github.com/jagdishlamba

## Misc

- Participated in National level 5G/6G hackathon in Oct 2024 by DoT and achieved 4th position.