

DEEPESH GUPTA

✉ hello@deepeshgupta.dev | 📞 +91 9929-285-466 | 🌐 deepeshgupta.dev | 🔗 LinkedIn | 🏠 Github

PROFESSIONAL SUMMARY

AI/ML Engineer specializing in Generative AI, LLM applications, and Computer Vision. Hands-on experience building production-ready RAG systems, document intelligence tools, and object detection pipelines. Proficient in LLM fine-tuning, YOLO model customization, prompt engineering, and deploying ML models with modern MLOps practices.

EDUCATION

B.Tech in Computer Science (AI & ML)

Oct 2021 – Jun 2025

KK Modi University

GPA: 8.72/10

- Relevant Coursework: Machine Learning, Deep Learning, NLP, Computer Vision, Data Structures & Algorithms

TECHNICAL SKILLS

Languages & Frameworks: Python, SQL, PyTorch, TensorFlow/Keras, Scikit-learn, XGBoost, FastAPI

Generative AI & LLMs: RAG Systems, LLM Fine-tuning, Prompt Engineering, LangChain, LlamaIndex, Hugging Face

Computer Vision & Vector DBs: OpenCV, YOLO, FAISS, Pinecone, Chroma, Semantic Chunking

MLOps & Cloud: Docker, Kubernetes, MLflow, GCP (Vertex AI, Cloud Run)

Data & Visualization: Pandas, NumPy, Matplotlib, Seaborn, Plotly

PROFESSIONAL EXPERIENCE

AI Research Engineer Intern

Feb 2026 – Present

Magure India Pvt. Ltd.

- Developing end-to-end object detection pipelines and fine-tuning YOLO models on custom datasets to deliver high-accuracy, production-ready computer vision solutions.

Localization Engineering Intern

Jul 2024 – May 2025

Andovar

- Automated file processing workflows using Python, reducing manual processing time by 40% for JSON, XML, TMX, PPTX, and DOCX formats.
- Optimized localization pipelines, improving project delivery efficiency and accuracy across multi-format document handling.

KEY PROJECTS

ResearchDigest – AI Research Analysis Platform | *Github*

- Built a full-stack app for natural language search and AI-powered summarization of research papers using HuggingFace transformers with semantic search and automated insight extraction.

MultiPDF ChatBot – Document Intelligence System | *Github*

- Developed a LangChain-based RAG system for multi-document Q&A with FAISS vector store and semantic chunking for accurate context retrieval across PDF documents.

ComputerUse – AI Computer Control Agent | *Github*

- Built an autonomous AI agent using vision models and LLM reasoning to automate UI interactions through screen understanding and intelligent action planning.

QuickTrack – Real-Time Object Detection System | *Github*

- Created a Streamlit app for real-time YOLOv8 object detection and tracking on video streams with multi-model support and downloadable annotated outputs.

CERTIFICATIONS

Oracle Cloud Infrastructure Data Science Professional · TensorFlow-Keras Bootcamp · Complete AI & ML Bootcamp · Elements of AI (University of Helsinki)