

Khadim Hussain

Islamabad, Pakistan | khadimxai@gmail.com | +92 311-6718-754

[in linkedin](#) | [upwork](#) | [github](#) | [huggingface](#)

SUMMARY

Applied ML Engineer with 4+ years building production computer vision and NLP systems for Industry 4.0. Delivered 70% inference speedups via TensorRT/Triton and improved model accuracy from 70% to 99.999% at elunic AG. Published fine-tuned LLMs and diffusion models on HuggingFace with 1.7k+ downloads.

EDUCATION

Bachelor of Science - Information Technology

University of Education

Sep 2020 - Apr 2024

Lahore, Pakistan

Master of Science - Computer Science

University of Southern Punjab

Oct 2024 - Sep 2026

Multan, Pakistan

SKILLS

Programming Languages	Python JavaScript TypeScript
Machine Learning	PyTorch TensorFlow Numpy OpenCV YOLO Computer Vision Defect Detection Deep Learning NLP LLM Transformers Unsloth PEFT Ollama
Technologies	Docker Kubernetes Git CI/CD CUDA Nvidia GPUs Flask FastAPI Hugging Face MLflow W&B MLOps Distributed Computing
Cloud Platforms	Azure Cloud GCP Vertex AI AWS Oracle
Databases	SQL MySQL Postgres Vector Databases Qdrant Pinecone Weaviate
Advanced Skills	Model Fine-tuning RAG Agentic AI LangChain LangGraph Prompt Engineering Multimodal AI vLLM RLHF LoRA Generative AI Diffusion Models Model Evaluation Red-teaming Synthetic Data Generation Cursor AI IDE Claude Code OpenAI Codex

EXPERIENCE

elunic AG

Senior Machine Learning Engineer

Jul 2022 - Present

Jan 2025 - Present

- Architected and built internal inference engines and Python tooling used across the ML team, optimizing model serving through TensorRT, ONNX, and Triton, resulting in a 70% improvement in inference speed.
- Led deep learning research initiatives and developed innovative engines for production deployment across multiple providers.
- Mentored and onboarded new team members, establishing training workflows and best practices for the ML team.

Data Scientist

Jul 2022 - Dec 2024

- Developed and fine-tuned segmentation and detection models using PyTorch and TensorFlow, deploying them with Docker and Kubernetes for efficient model management in an MLOps environment.
- Achieved an accuracy rate of 99.999% in most cases, significantly improving company model performance from 70% to 99.999%, a gain of 29.999%.
- Led data annotation efforts and managed a team to ensure high-quality annotated datasets for model training.

Freelance (Upwork)

LLM Developer

Feb 2022 - Dec 2023

Toronto, Ontario

- Collected and annotated an 11K-sample text dataset for an LLM project, implementing models for text classification to generate answers based on questions and context.
- Developed efficient data collection and annotation pipelines to streamline dataset preparation for LLM training.

Freelance (Upwork)

Computer Vision Engineer/Data Scientist

Jan 2022 - Dec 2023

New York, NY

- Built end-to-end TensorFlow models for computer vision and LLM, optimizing training pipelines and conducting performance evaluations to iteratively refine models, ensuring they meet key metrics and deliver optimal results.
- Engineered inference pipelines, achieving 90% faster predictions and enhancing deployment efficiency.

NOTABLE PROJECTS

Kiffusion & SpaceVision

Dataset Creation & Diffusion Model Fine-tuning

Jan 2025 - Feb 2025

[HuggingFace](#)

- Published a 1,964 image-caption dataset (1.7k+ downloads on HuggingFace) and trained LoRA adapters on FLUX.2-klein (9B) and Kolors for photorealistic space image generation using int8 quantization via SimpleTuner.

STEM Reasoning LLMs

LLM Fine-tuning & Quantization

Jan 2025 - Feb 2025

[HuggingFace](#)

- Fine-tuned GPT-OSS-20B and Qwen3-14B for STEM reasoning and Q&A using QLoRA and Unsloth. Published LoRA, merged, and GGUF-quantized versions compatible with Ollama, llama.cpp, and LM Studio.

LANGUAGES

- English - Fluent