DEEP PATEL

pateldeep1842@gmail.com | LinkedIn | GitHub
Seattle, WA, USA | Contact: +1 (425) 305-1382

SUMMARY

AI/ML Engineer experienced in fine-tuning large language models (GPT, LLaMA) and building scalable NLP pipelines using Hugging Face, PyTorch, and RAG architectures. Proficient in deploying production-ready AI solutions with FastAPI, Docker, and AWS Lambda. Applied Generative AI, FAISS-based vector search, and reinforcement learning in real-world enterprise and academic projects.

EDUCATION

MS in Computer Science — Aug 2023 - May 2025
University of Oklahoma, Norman, OK, USA

BTech in Computer Science and Engineering — Aug 2018 – May 2022
Indian Institute of Information Technology (IIIT), Gujarat, India

TECHNICAL SKILLS

LLMs & NLP: GPT-2, LLaMA, Mistral, Hugging Face Transformers, PEFT, LangChain, FAISS, RAG, LangGraph.
ML & Deployment: PyTorch, Scikit-learn, TensorFlow, Reinforcement Learning, FastAPI, Docker, AWS Lambda, Firebase, CI/CD.
Programming & Tools: Python, Java, SQL, NumPy, Pandas, OpenCV, Jupyter, Git, Linux CLI, MATLAB, Plotly, Matplotlib, LaTeX.

EXPERIENCE

AI Engineer at Firenix Technologies Pvt. Ltd. — May 2022 – May 2023

- Developed end-to-end facial recognition and anomaly detection systems using PyTorch and OpenCV; reduced false access events by 28% in smart surveillance setups.
- Applied model compression and 4-bit quantization for real-time edge deployment, improving inference speed by 40%; methods adapted later for lightweight LLM inference.

Machine Learning Research Intern at IIIT Vadodara — May 2021 – Jul 2021

- Designed and trained a CNN on 492 PET scans for Alzheimer's detection, achieving 82% accuracy and outperforming state-of-the-art baselines.
- Improved model robustness with neuroimaging-specific preprocessing and data augmentation; validated performance through literature review and benchmarking.

PROJECTS

NeuroDoc: Multi-Document RAG Assistant with Citation and Memory `github` — Feb 2025

- Developed a multi-document RAG system (FAISS + BM25) enabling PDF upload and context-aware querying; achieved 200ms average retrieval latency and 18% boost in answer relevance via reranking and citation-grounded responses.
- Integrated session-based memory and chat logging to support follow-up queries and long-term usage analysis, enhancing contextual understanding and model evaluation.

VisuaLens: Multimodal LLM System `github` — Nov 2024

- Engineered a multimodal LLM application using LLaVA to support dual-image upload and comparative question answering for tasks like chart interpretation and UI analysis via grounded visual reasoning.
- Integrated prompt templates, image preprocessing, and response logging to enhance visual-text alignment and reduce hallucination; optimized inference using 4-bit quantized models for local deployment.

LiteLLM: Local LLM Deployment — Mar 2024

- Built a fully local LLM inference pipeline using LLaMA with 4-bit quantization (bitsandbytes) and LoRA fine-tuning, reducing GPU memory usage by 60% and improving response speed by 40%, enabling efficient offline text generation.

CERTIFICATES

Generative AI with Large Language Models by deeplearning.ai
Natural Language Processing with Transformers by Hugging Face
Building Transformer-Based Natural Language Processing Applications by Nvidia Deep Learning Institute
Fundamentals of Deep Learning by Nvidia