DEEP PATEL

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

SUMMARY

Results-driven AI/ML Engineer specializing in fine-tuning large language models (GPT, LLaMA, Mistral) and architecting scalable NLP solutions. Experienced in building end-to-end systems with Retrieval-Augmented Generation (RAG), FAISS vector search, and Hugging Face Transformers. Proficient in MLOps, deploying production-ready models using Docker, FastAPI, and AWS, with a strong foundation in PyTorch and reinforcement learning.

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 & Generative AI: LLaMA, GPT-2/3, Mistral, RAG, LangChain, LangGraph, Hugging Face (Transformers, PEFT), FAISS, Vector Databases, Prompt Engineering, Fine-Tuning (LoRA).
MLOps & Deployment: PyTorch, TensorFlow, Scikit-learn, Docker, FastAPI, AWS (Lambda, S3), Firebase, CI/CD, Git, Linux CLI, Model Quantization, Inference Optimization.
Core Programming & Data Science: Python, Java, SQL, NumPy, Pandas, OpenCV, Jupyter, MATLAB, Plotly, Matplotlib, LaTeX.

EXPERIENCE

AI Engineer at Firenix Technologies Pvt. Ltd. — May 2022 – May 2023
- Engineered and deployed an end-to-end facial recognition system using PyTorch and OpenCV, reducing false access events in smart surveillance setups by 28%.
- Quantized production models to 4-bits and applied compression techniques, improving inference speed by 40% for real-time edge deployment.

Machine Learning Research Intern at IIIT Vadodara — May 2021 – Jul 2021
- Trained and benchmarked a custom CNN on 492 PET scans for Alzheimer's detection, attaining 82% accuracy and outperforming existing state-of-the-art models.
- Enhanced model robustness by implementing neuroimaging-specific data augmentation and preprocessing pipelines to ensure generalizability and clinical relevance.

PROJECTS

NeuroDoc: Multi-Document RAG Assistant with Citation and Memory `github` — Feb 2025
- Architected a multi-document RAG system with a hybrid search retriever (FAISS vector search + BM25) to provide citation-grounded answers from uploaded PDFs, boosting relevance by 18%.
- Integrated stateful conversational memory and chat logging, allowing the system to handle follow-up questions and maintain context across user sessions.

VisuaLens: Multimodal LLM System `github` — Nov 2024
- Built a multimodal question-answering application using LLaVA for comparative visual analysis that lets users upload multiple images and query their content.
- Refined visual-text alignment and mitigated model hallucination through advanced prompt engineering; employed 4-bit quantization for effective local deployment.

LiteLLM: Local LLM Deployment — Mar 2024
- Developed a lightweight, fully local inference pipeline by fine-tuning a LLaMA model with LoRA; leveraged 4-bit quantization to slash GPU memory usage by 60% for 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