DEEP PATEL

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

SUMMARY

AI/ML Engineer with expertise in supervised & unsupervised learning (XGBoost, Random Forests, Clustering) and deep learning (CNNs, RNNs, Transformers). Skilled in fine-tuning LLMs (GPT, LLaMA, Mistral with LoRA/QLoRA) and deploying scalable AI systems using Docker, FastAPI, AWS, and Kubernetes.

TECHNICAL SKILLS

ML: Supervised & Unsupervised Learning, Regression, Classification (Random Forest, XGBoost), Clustering (K-Means), Feature Engineering, Model Evaluation.

NLP & LLMs: Hugging Face Transformers, GPT, LLaMA, Mistral, LoRA/QLoRA, RAG Pipelines, LangChain, LangGraph, Prompt Engineering, Vector Databases (FAISS, Pinecone, Qdrant).

MLOps & Deployment: PyTorch, TensorFlow, Docker, FastAPI, AWS (Lambda, S3, EC2), Kubernetes, CI/CD, MLflow, Weights & Biases.

Programming & Tools: Python, Java, SQL, NumPy, Pandas, OpenCV, Matplotlib, Git, Linux CLI.

PROJECTS

NoteBook LLM: Multi-Document RAG Assistant github — Jun 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.

Efficient Fine-Tuning of LLMs with LoRA & QLoRA github — Feb 2025

- Reduced trainable parameters by 95% and cut training time up to 10× while achieving 12% lower perplexity on GPT-2 and comparable accuracy on BERT (52.17%) using LoRA and QLoRA fine-tuning.

Customer Segmentation & Prediction github — Nov 2024

- Implemented RFM-based customer segmentation with K-Means clustering (Elbow method, K=4) to identify high-value vs. low-engagement customers, improving business insight through cluster analysis and visualizations.
- Built and tuned an XGBoost classifier with GridSearchCV to predict next-month purchases (77% accuracy, F1=0.73) and a Linear Regression model to estimate customer spend (MAE ≈ $275), enabling actionable retention and revenue strategies.

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.

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

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