DEEP PATEL

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

SUMMARY

AI/ML Engineer with 4+ years of experience specializing in LLMs, RAG systems, and end-to-end ML pipelines. Developed production-grade retrieval, multimodal reasoning, and fine-tuning workflows using PyTorch, Transformers, and AWS. Strong in MLOps, scalable backend engineering, and deploying AI systems at scale.

TECHNICAL SKILLS

Programming & Tools: Python (3.x), Java, SQL, NumPy, Pandas, OpenCV, Matplotlib, Git, Linux (CLI), OOP, Data Structures & Algorithms, RESTful APIs (FastAPI, Flask)

Data & Databases: SQL, PostgreSQL, MySQL, DynamoDB, Pandas, NumPy, ETL Pipelines, Data Modeling

Machine Learning & Deep Learning: Supervised & Unsupervised Learning, Regression, Classification (Random Forest, XGBoost), Clustering (K-Means), Dimensionality Reduction (PCA), Feature Engineering, Model Evaluation, PyTorch, TensorFlow

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

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

PROFESSIONAL EXPERIENCE

Community Dreams Foundation — *Remote, USA*
Machine Learning Engineer — Aug 2025 – Present

- Conducted end-to-end machine learning lifecycle activities, including requirements gathering, model design, training, and evaluation, improving predictive model performance by 22%.
- Built and maintained data pipelines and training datasets using Python and SQL, enabling consistent data ingestion, preprocessing, and reporting for ongoing model monitoring.
- Analyzed large community-impact datasets to identify trends and patterns, generating actionable insights and supporting process improvements that increased model reliability by 30%.

TripRaft — *Remote, USA*
Founding Machine Learning Engineer — May 2025 – Aug 2025

- Architected the core platform and delivered scalable backend services for itinerary planning, optimizing database queries and caching to reduce request latency by 30%.
- Engineered secure data schemas and RESTful APIs for complex interactions such as collaborative voting and expense sharing, ensuring data integrity and reliable performance under concurrent usage.
- Delivered ML-driven personalization modules including destination ranking and semantic search using vector embeddings improving trip discovery relevance by 18%.

Firenix Technologies — *India*
AI/ML Engineer — Mar 2021 – Jul 2023

- Developed and deployed high-accuracy PyTorch models for real-time classification tasks, achieving a 28% reduction in prediction errors post-deployment.
- Optimized models for edge computing using 8-bit quantization and ONNX Runtime, cutting inference latency by 40% while maintaining model accuracy.
- Designed and orchestrated end-to-end ML pipelines covering feature engineering, data processing, and API integration with FastAPI and Docker ensuring robust deployment across multiple client projects.

IIIT Vadodara — *India*
ML Research Intern — Aug 2020 – Dec 2020

- 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

NoteBook LLM – Multi-Document RAG Assistant github Jan 2025

- Constructed a multi-document RAG pipeline using hybrid retrieval (FAISS vector search + BM25) to deliver citation-grounded responses across long PDF inputs.
- Integrated session-aware conversational memory that preserved user context, enabling accurate follow-up queries across multi-turn interactions.
- Enhanced retrieval precision by 18% through strategic reranking, metadata normalization, and structured prompting techniques.

VisuaLens – Multimodal LLM System github Nov 2024

- Constructed a multimodal visual–text reasoning framework using LLaVA, enabling image comparison, description, and analytical Q&A on user-uploaded media.
- Refined visual–language alignment through targeted prompt engineering and consistency checks, significantly reducing hallucination rates.
- Accelerated local inference with 4-bit quantization, enabling smooth on-device processing without relying on cloud infrastructure.

Spendly – Smart Expense Manager Feb 2024

- Established secure authentication, group expense workflows, and automated settlement logic to support accurate and transparent multi-user financial tracking.
- Designed a hybrid data storage architecture using PostgreSQL for transactions and DynamoDB for session/state management, ensuring scale across 10,000+ records.
- Implemented interactive data visualizations with Plotly and Pandas, reducing manual expense analysis time by 75% and improving user reporting workflows.

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