DEEP PATEL

Deep.Patel-1@ou.edu | LinkedIn | GitHub
Norman, OK, USA | Contact: (+1) 4253051382

SUMMARY

Experienced in LLM fine-tuning (GPT, Llama), RAG architectures, and NLP workflows using Hugging Face Transformers and spaCy. Skilled in building end-to-end ML pipelines, statistical modeling, and deploying scalable applications. Proficient in Python and PyTorch with a focus on optimizing model performance, data preprocessing, and translating research into production-ready AI solutions. Collaborative developer passionate about solving complex problems with efficient algorithms and modular software design.

EDUCATION

MS in Computer Science — Expected Graduation, May 2025
University of Oklahoma (OU) — OK, USA

BTech in Computer Science and Engineering — 2018-2022
Indian Institute of Information Technology (IIIT) — GJ, India

SKILLS & COURSEWORK

Programming Languages: Python, Java, SQL, PHP, HTML, MATLAB

Machine Learning: PyTorch, TensorFlow, Scikit-learn, Neural Networks, Reinforcement Learning

Generative AI: LLM Fine-Tuning, RAG, Vector Databases (FAISS), Prompt Engineering, Agentic AI

NLP: spaCy, Hugging Face Transformers

Cloud & MLOps: AWS, Docker, FastAPI, CI/CD Pipelines, Azure AI, Google Firebase

Data Tools: Pandas, NumPy, OpenCV, Jupyter Notebook

Relevant Coursework: Data Structures & Algorithms, Database Management Systems,Machine Learning, Artificial Intelligence,Natural Language Processing (NLP), Software Engineering, PDN Programming, Network Science

AI/ML PROJECTS

EcoAssist: RAG-Powered E-Commerce Chatbot — Jan 2025 - Present
Ongoing Project

- Developed a customer support chatbot using LLMs (GPT, Llama) fine-tuned on 5,000+ domain-specific interactions via Hugging Face, optimizing responses for order-tracking and product inquiries.
- Implemented RAG architecture with FAISS vector database to enhance answer accuracy through real-time data retrieval.
- Deployed the solution on AWS Lambda using FastAPI and Docker, ensuring scalability for enterprise-level traffic.

ScoreVision: IPL Score Prediction System — Nov 2024
Network Science

- Built a predictive engine with Scikit-learn/XGBoost on IPL data (2008–2024), engineering 50+ features (player stats, pitch conditions) to forecast match outcomes and target scores.
- Incorporated Gemini API for real-time commentary and deployed Llama-2 70B to synthesize post-match reports.
- Automated data preprocessing workflows for IPL datasets, reducing manual cleaning time by 40%, and visualized insights with interactive dashboards to highlight player performance trends and strategic recommendations.

JetNav: Autonomous Navigation with Nvidia JetBot — Mar 2024
Course Project

- Engineered a ResNet-18 model for real-time road detection using PyTorch and OpenCV, achieving collision avoidance accuracy in the range of 80% in dynamic environments (e.g., cluttered indoor spaces)
- Integrated SLAM algorithms to enable autonomous navigation, reducing manual intervention by 30% compared to baseline rule-based systems.
- Optimized inference speed by 15% using PyTorch's quantization tools, deploying the model on AWS EC2 via Docker for edge computing.

SOFTWARE ENGINEERING PROJECTS

EzyShop: Responsive ECommerce Site May 2023

- Crafted a full-stack platform with Python, SQL, and HTML/CSS, featuring user authentication, cart management, and order tracking for a seamless e-commerce experience.
- Enhanced database efficiency by 25% through strategic indexing and caching, supporting smooth scalability for 100+ product listings
- Established REST API integrations for payment gateways and shipping services, trimming third-party service setup time by 30% in development.

ScanMaster: Document Management App Dec 2022

- Designed a document scanner app with Python and OpenCV, enabling real-time image preprocessing (cropping, noise reduction) for 500+ scanned pages during testing.
- Automated text extraction using optical character recognition (OCR) and spaCy's NLP pipeline, reducing manual data entry time by 60% for structured documents
- Containerized the application with Docker and deployed it on AWS EC2, integrating Firebase for secure cloud storage and user authentication.

WORK EXPERIENCE

IIIT Vadodara May 2021 - Jul 2021
Machine Learning Research Internship

- Architected and trained a Convolutional Neural Network (CNN) model to detect Alzheimer's disease using PET scans from 492 patients, achieving 82% accuracy and surpassing state-of-the-art deep learning architectures.
- Processed and analyzed large-scale medical imaging data using Python and machine learning frameworks, enhancing model performance through advanced data augmentation techniques.

MentorBoxx Dec 2020 - Jan 2021
UI/UX Designer

- Engineered an Android Travel Guide app using Dart and Node.js, integrating Google authentication via AWS for secure user sign-in and seamless access.
- Shaped and streamlined user interfaces using Figma, increasing user engagement by 35% through iterative usability testing and responsive design.
- Collaborated with a cross-functional team to deliver an intuitive and user-friendly application, ensuring timely project completion and high user satisfaction.

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

Machine Learning by Stanford University on Coursera

Deep Learning by IBM