Nainesh Harshad Rathod

Palo Alto, CA | nainesh.h.rathod@gmail.com | (812) 671-3542 | LinkedIn | GitHub

Summary

Data Science Master's graduate and AI Engineer with expertise in productionizing LLM solutions, including RAG pipelines (Azure, LangChain) and fine-tuning Llama-3 for context-aware Q&A. Seeking to apply Python, PyTorch, and cloud engineering skills to solve Gen-AI challenges—delivering models, not just prototypes.

Education

Indiana University Bloomington | *MS in Data Science* | *GPA: 3.7/4* — Aug 2023 – May 2025
Coursework: Machine Learning, Statistical Inferencing, Advanced Database concepts, Data Mining, Time series analysis

University of Mumbai | *BE in Computer Engineering* | *GPA: 8.9/10* — Aug 2019 – May 2023

Technical Skills

Programming Languages: Python, R, SQL, C++
ML & AI Frameworks: Pandas, NumPy, Scikit-learn, TensorFlow, PyTorch, Keras, OpenAI API, Hugging Face
Generative AI & LLMs: Prompt engineering, Fine-tuning LLMs
Cloud & Deployment: AWS, Azure AI Services, Docker, Kubernetes
Data Engineering & Analytics: Data cleaning and preprocessing, Feature engineering, Data visualization (Matplotlib, Seaborn, Plotly), A/B testing

Work Experience

Machine Learning Engineer — Dec 2024 – present
Kelley School of Business @ Indiana University — *Bloomington, IN*

- Architected and implemented an end-to-end Retrieval Augmented Generation (RAG) pipeline on Azure for complex legal documents, enabling semantic querying of 1,000-page publications.
- Integrated and leveraged an internally hosted Llama 3.1 (8B) within a secure Azure environment, applying it for context-aware Q&A, with average query response times under 5 seconds.
- Engineered automated data processing workflows for 1,000-page PDFs (parsing, chunking, embedding) feeding a vector database of over 100,000 chunks, ensuring efficient data ingestion for LLM applications.

Machine Learning Engineer Intern — Jul 2024 – Nov 2024
Hyphenova – a tech startup connecting influencers and brands on social media – [Company info] — *Los Angeles, CA*

- Fine-tuned BERT (Hugging Face) on 50K song lyrics for sentiment analysis, achieving 85% accuracy (15% gain vs. rule-based baselines), enabling data-driven music trend analysis.
- Fine-tuned and deployed advanced computer vision models like YOLOv11, ResNet, and Vision Transformers to detect and segregate sensitive content on social media platforms, achieving an overall accuracy of 93%.
- Developed a Python-based web scraper to collect and analyze 100,000+ Instagram posts, extracting user engagement metrics (likes, comments, hashtags) and enabling insights into content performance trends.

Machine Learning Engineer Intern — Feb 2022 – Apr 2023
Sanskritech – a healthcare startup building point-of-care medical solutions - [Company info] — *Mumbai, India*

- Spearheaded the development of an end-to-end ML pipeline for a fertility analyzer, by model fine-tuning using YOLOv5, PyTorch, and OpenCV, achieving 85% validation accuracy.
- Built an NLP pipeline with spaCy to perform sentiment analysis (88% accuracy) of 10K+ customer reviews, leveraging tokenization, lemmatization, and custom ML models (textcat), improving feedback processing speed by 90%.
- Slashed ticket triage time by 60% via a TF-IDF + SVM classifier, prioritizing critical bugs and reducing customer escalation rates by 25%.

Projects

Contextual Document Chatbot with LLM | *LLM, Langchain, OpenAI GPT, Pinecone, FAISS* – GitHub link

- Developed a document-based Question-Answering (QA) system using LangChain and GPT 3.5, achieving an average retrieval speed of 25ms with FAISS, enabling real-time querying of a document corpus.
- Engineered a Retrieval-Augmented Generation (RAG) pipeline that achieved 92% top-5 retrieval accuracy through strategic fine-tuning of document chunking and embedding selection, ensuring highly relevant context for LLM responses and showcasing large model application.

Real-time Traffic Sign Classification | *PyTorch, Deep Learning, Computer Vision* – GitHub link

- Developed and deployed a real-time traffic sign recognition system using the GTSRB dataset, achieving 98% test accuracy across 43 classes, demonstrating practical application of deep learning for real-world problems.
- Fine-tuned a pre-trained MobileNetV2 model and deployed an interactive web application using Streamlit, enabling real-time image uploads and predictions.

Diabetes Prediction using Machine Learning | *Classification, SMOTE, EDA, Data Analysis* – GitHub link

- Developed a diabetes risk prediction model from 253,000+ survey responses, enabling healthcare providers to proactively identify high-risk patients and recommend early intervention strategies.
- Implemented ML algorithms: SVM, Random Forest and Decision Tree, achieving a peak accuracy of 86% and a recall of 96%, addressing missing values, feature engineering, and balancing classes using SMOTE.

MediTrack – Hospital Inventory Management System | *MERN Stack* – GitHub link

- Architected and implemented a full-stack hospital inventory management system using the MERN stack (MongoDB, Express.js, React.js, Node.js).
- Designed modular RESTful APIs for CRUD operations, integrated secure role-based access, and implemented robust data models to handle 200+ real-world inventory items.