Dipen Dave

AI Engineer
+91-799-010-1442 • dipendave.ai@gmail.com • linkedin.com • Ahmedabad, Gujarat, India

Summary

AI/ML Engineer with hands-on experience in building end-to-end machine learning and AI applications using Python, Fast API, and modern ML frameworks. Skilled in developing scalable pipelines, retrieval-based systems, and backend solutions with PostgreSQL and AWS services. Passionate about solving real-world problems, learning new technologies quickly, and delivering efficient, production-ready solutions. Eager to contribute to innovative AI/ML projects while expanding technical expertise.

Education

Gujarat Technological University — Ahmedabad, Gujarat, India
Bachelor of Technology in Computer Science with (AI/ML) | GPA: **9.1**/ 10 — 06/2020 - 06/2024

- Worked on machine learning projects using different algorithms to make predictions from sample data.
- Built a solid understanding of statistical reasoning and probability theory to support the creation of accurate and reliable ML models.

Experience

The Intellify — Ahmedabad, India
AI/ML Engineer (07/2025 - Present) / AI/ML Engineer Intern (01/2025 – 06/2025)

- Led end-to-end development of a **serverless ML system** to predict sports betting odds, deployed via **Lambda-compatible Docker containers**; handled full project lifecycle independently in the absence of senior engineers.
- Architected a modular, containerized ML pipeline using **Python**, **Scikit-learn**, and **Lambda-compatible Docker**, enabling scalable, low-latency predictions in a fully serverless **AWS Lambda** environment. Improved model performance by ~**25%** through advanced **feature selection**, **L1/L2 regularization**, and **hyperparameter tuning**, selecting the optimal algorithm via rigorous **model benchmarking**.
- **Developed an end-to-end RAG-based document chatbot** with real-time streaming using **Fast API** and **WebSocket**, featuring **authentication**, **PostgreSQL (pgvector)**, and seamless **Lang Chain–OpenAI** integration.
- Engineered a scalable, dynamic retrieval pipeline with optimized **embeddings**, enhancing response accuracy by **15%** and UX for an in-house AI product deployed on **AWS EC2**.
- Developed an **Agentic AI-driven POC** for a healthcare insurer using **Lang Graph** and intelligent API orchestration to deliver real-time, ID-based plan recommendations and dynamic coverage insights—designed, built, and deployed end-to-end in under **1 week**.
- Built an AI agent with Crew AI to generate Google queries from job criteria, extract LinkedIn profiles, and structure data into clean datasets, making candidate sourcing faster and more reliable.
- Developed an automated resume-screening pipeline where incoming resumes were stored in S3, parsed via AWS Lambda to extract candidate details, and filtered with an Agentic AI that generated SQL queries and fetch the best-fit candidates from the database according to job description given by the user which reduce manual hiring effort by ~20% and accelerating candidate shortlisting.

IGENERATE Technologies. — Ahmedabad,Gujarat, India
AI/ML Engineer Intern — 04/2024 - 10/2024

- Developed data scraping scripts using various python-based data scraping libraries such as Beautifulsoup4, scrapy etc., for Scrap restaurant's menu data from online aggregators.
- Became proficient with Hugging Face tools and pretrained large language models (LLMs), fine-tuning a model on a custom dataset.
- Built recommendation systems and sentiment analysis models for real-time food industry data.
- Created End to End AI/LLM application with Stream lit and Fast API to develop APIs and deployed APIs on the Hugging Face server.
- Explored Lang Chain and Llama Index, using vector-based embeddings create RAG (Retrieval-Augmented Gen eration) agents for retrieving data from unstructured data such as PDFs, Articles etc.

Projects

Document Query System with LLM-based RAG

- Built a full-stack intelligent Q&A system over PDFs using LLMs and Retrieval-Augmented Generation with real-time streaming responses.
- Integrated pgvector for vector storage, Ollama Embeddings for chunking, and connected a Stream lit frontend to a Fast API backend.
- Used Docker Compose for scalable deployment and implemented asynchronous query handling for improved performance

E-commerce Backend with Microservices

- Built a modular backend for an e-commerce platform using **Fast API**, implementing **JWT-based authentication and encrypted APIs** with separate services for products, orders, and payments to ensure security and scalability.
- Designed relational database schemas in **PostgreSQL** to manage users, inventory, and transactions with optimized queries for faster access.
- Containerized the system with Docker for consistent and scalable multi-environment deployment.

Skills

Programming Languages: Python

Web Frameworks: Lang chain • Lllama-Index • Fast API • Flask

Data Science & Machine Learning: Pandas NumPy • Scikit-learn • TensorFlow • PyTorch

Databases: MySQL • PostgreSQL • MongoDB • Vector Databases

Data Visualization: Tableau • Matplotlib • Seaborn

Familiar with: AWS EC2 • AWS S3 • AWS Lambda

Soft Skills: Problem-Solving • Critical Thinking • Teamwork • Communication • Interpersonal Skills • Quick Learner