Dhruv Mendiratta

dhruv.mendiratta4@gmail.com | +919013669130 | dhruv-portfolio-bay.vercel.app | linkedin.com/in/dhruvmendiratta18 | github.com/dhruvm-18

SUMMARY

Final-year Computer Science Engineering student specializing in Artificial Intelligence Engineering, with hands-on experience in Generative AI, Retrieval-Augmented Generation (RAG) systems, embedding optimization, and vector search. Three-time published AI researcher with expertise in Python, PyTorch, TensorFlow, FastAPI, React, and scalable AI product enhancement. Proven ability to increase model accuracy by up to 30% and reduce query latency by 20% through optimized pipelines and embeddings. Skilled in transformer-based architectures, cloud-based AI deployments (AWS), and full-stack AI solutions.

SKILLS

Programming: Python, C, SQL, JavaScript
Big Data: Hadoop, Sqoop, Hive, Kafka
AI/ML: LangChain, Large Language Models (LLMs), FAISS, TensorFlow, Keras, Scikit-learn, PyTorch, Hugging Face, Transformer Models, Prompt Engineering, Model Deployment
Data Tools: Pandas, NumPy, Matplotlib, GARCH, Data Preprocessing, Exploratory Data Analysis (EDA), Tableau, Power BI Data Management, Data Governance
Web & APIs: FastAPI, React.js, Node.js, REST APIs, PostgreSQL, Flask
Consulting & Professional Skills: Documentation, Requirements Gathering, Client Communication, Business Process Analysis, Solution Design, Stakeholder Management
Cloud: AWS (EC2, S3, Lambda, IAM, CloudWatch), Git
Other Tools: MongoDB, JIRA, NLTK, Chart.js, Framer Motion, Leaflet.js

EDUCATION

Manipal University Jaipur — Jaipur, Rajasthan
B.Tech in Computer Science and Engineering — Sept 2022 – Present
- Awarded Student Excellence Award (×2) — August 2024 and March 2025 — for excellence in internship performance and AI research publications; authored three peer-reviewed AI papers published in international conferences and journals.

WORK EXPERIENCE

Ernst & Young (EY) — Delhi, India
AI Intern, Generative AI & RAG Systems — May 2025 – July 2025
- Interned as an AI/ML Engineer, developing optimized retrieval and embedding pipelines that improved precision by 30% and reduced latency by 20%.
- Fine-tuned transformer models and built low-latency inference workflows to deliver accurate, real-time chatbot responses.
- Integrated hallucination mitigation and compliance-focused prompt workflows, cutting manual knowledge lookup time by 50% and ensuring factual accuracy.

Deloitte Touche Tohmatsu LLP — Gurugram, Haryana
Software Engineering Intern, Cloud Computing (AWS) — May 2024 – July 2024
- Gained hands-on experience in AWS services including EC2, S3, Lambda, IAM, and CloudWatch, applying them in cloud deployment scenarios.
- Practiced real-world deployment scenarios involving cloud architecture, scalability, and security.
- Collaborated in infrastructure review and optimization sessions, improving team understanding of cloud monitoring, resource utilization, and security controls.
- Contributed to scalability improvements in AWS cloud deployments.

PROJECT WORK

Unified Knowledge Platform — Enterprise RAG Chatbot — May 2025 – July 2025
- Tech Stack: LangChain, FAISS, Gemini API, Python
- Designed and deployed an enterprise document Q&A chatbot using optimized embeddings, FAISS vector search, and fine-tuned prompt workflows, improving retrieval precision by 30%, reducing query latency by 20%, and enhancing enterprise knowledge access.

CrisisReport — Crowd-Sourced Disaster Reporter — May 2025 – July 2025
- Tech Stack: React.js, Flask, MUI, Leaflet.js, Axios, Framer Motion, REST APIs, Flask-CORS
- Developed a full-stack disaster reporting platform with real-time incident uploads, interactive geolocation maps (Leaflet.js + Flask APIs), and orchestrated REST API integrations, accelerating emergency response coordination.

Product Sentiment Analyzer (Full-Stack NLP App) — March 2025
- Tech Stack: FastAPI, React, NLTK, PostgreSQL
- Built a real-time sentiment analysis system with a custom NLP pipeline, interactive React visualizations, and PostgreSQL-backed storage, improving classification accuracy and enabling faster product team decision-making.

PUBLICATIONS

CML 2025: Skin Disease Detection using CNN+GAN Models on ISIC Dataset (96.3% accuracy).
ICAESRTA 2025: Hybrid BiLSTM-GRU Model for S&P 500 Forecasting ($R^2 = 0.9901$).
Cuestiones de Fisioterapia (2025): Role of AI in Radiology and Medical Diagnostics.

HONORS & AWARDS

Student Excellence Award (×2): August 2024 and March 2025 for excellence in internship performance and AI research publications.
HackX Finalist (2024): Reached final round of Manipal University's premier hackathon.
Conference Presentations: Presented research at CML 2025 (Skin Disease Detection using CNN+GAN Models) and ICAESRTA 2025 (Hybrid BiLSTM-GRU Model for Stock Forecasting).