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
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)
Web & APIs: FastAPI, React.js, Node.js, REST APIs, PostgreSQL, Flask
Cloud: AWS (EC2, S3, Lambda, IAM, CloudWatch), Docker, Kubernetes, 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

- Applied embedding optimization strategies to improve precision by 30% and reduce latency by 20%.
- Enhanced retrieval accuracy using transformer fine-tuning for domain-specific responses.
- Engineered low-latency inference workflows for real-time enterprise chatbot responses.
- Integrated hallucination mitigation techniques to ensure factual accuracy in AI responses.
- Developed optimized retrieval pipelines for large-scale document queries.
- Built scalable ingestion and chunking pipelines with optimized embeddings for real-time, high-precision queries.
- Integrated prompt workflows with proprietary document stores, reducing manual knowledge retrieval time by 50% and ensuring compliance in automated responses.

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 internal enterprise document Q&A chatbot, improving retrieval precision by 30% and reducing query latency by 20% through optimized embeddings and document chunking.
- Integrated FAISS vector search and fine-tuned prompt workflows for higher contextual accuracy in automated responses.
- Enhanced enterprise knowledge access by reducing manual lookup times and improving compliance in generated outputs.

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 enabling citizens to upload incidents in real-time, accelerating emergency response coordination.
- Integrated interactive geolocation maps using Leaflet.js and CORS-enabled Flask APIs for seamless reporting.
- Coordinated multiple REST API calls using API orchestration patterns for seamless data flow.

Product Sentiment Analyzer (Full-Stack NLP App) — March 2025

- Tech Stack: FastAPI, React, NLTK, PostgreSQL
- Created a real-time sentiment analysis application processing customer product reviews using a custom NLP pipeline, improving classification accuracy.
- Built an interactive React dashboard with dynamic sentiment visualizations to speed up decision-making for product teams.
- Developed a PostgreSQL-backed storage system for fast data retrieval and long-term review analytics.

Hybrid Stock Price Prediction Model — January 2025

- Tech Stack: LSTM, GARCH, TensorFlow, Keras, Python
- Engineered a hybrid deep learning model combining LSTM for sequential trend detection and GARCH for volatility modeling.
- Achieved $R^2 = 0.9901$ and RMSE = 0.0125 on S&P 500 forecasts, enabling highly accurate predictive insights.
- Published results in the ICAESRTA 2025 conference, validating performance through peer review.

Railway Ticketing Chatbot (Rule-Based NLP) — May 2025

- Tech Stack: Python, NLTK
- Developed a rule-based chatbot for train ticket booking simulation, incorporating keyword mapping and intent classification.
- Enhanced user experience by implementing input validation and automated query handling for realistic text-based booking flows.

PUBLICATIONS

CML 2025: Skin Disease Detection using CNN+GAN Models on ISIC Dataset: Achieved 96.3% accuracy with Grad-CAM for model interpretability.
ICAESRTA 2025: Hybrid BiLSTM-GRU Model for S&P 500 Forecasting: Attained $\mathbf{R^2} = 0.9901$ and MAE = 0.0101 using deep learning-based time-series analysis . Evaluated models using $\mathbf{R^2}$, RMSE, MAE, and other model evaluation metrics.
Cuestiones de Fisioterapia (2025): Role of AI in Radiology and Medical Diagnostics: Published peer-reviewed journal article highlighting AI's applications in medical imaging.

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).