Divyansh Pandey

Lucknow, UP, India | +91-9305425557 | divyanshpandey0108@gmail.com
LinkedIn | Portfolio | Github

Education

Manipal University Jaipur — Jaipur, Rajasthan
B.Tech (Hons.) Computer Science Engineering (AIML) | **CGPA**: 8.89 — Sept. 2022 – June 2026

Experience

AI/ML Engineer Intern — Hyderabad, Telangana, India
VIGIL-Labs (IIT-H) | Link — Apr. 2025 – July 2025

- **Accelerated** model training convergence by **optimizing** communication protocols, **reducing** global communication round **time by 45%** and utilizing **55% fewer** resources than baselines.
- Engineered a **decentralized** Federated Learning model for medical image classification, **surpassing baseline test accuracy by 20%** on complex non-IID **real-world** data.
- Orchestrated secure machine learning workflows and formulated **scalable** algorithmic improvements to address data heterogeneity, translating project requirements into a robust AI solution that ensured strict distributed data **privacy**.

Projects / Open-Source

Real-Time Fraud Detection System | Python, FastAPI, Docker, AWS | Repo

- **Accelerated** prediction **latency by 90% (50ms to 5.4ms)**, engineering a high-performance **FastAPI** microservice with **asynchronous** lifespan management to **surpass** strict SLAs for **real-time** payment processing.
- **Maximized** False Positive precision **from 6% to 78%** by evaluating a class-weighted XGBoost **challenger** model against baselines using MLflow, maintaining **83% Recall** to drastically reduce customer friction.
- Established **100% observability** into production model health by integrating custom **Prometheus** metrics and **Grafana** dashboards, protecting against financial risks from **concept drift** and silent failures.
- Constructed a resilient **Dockerized** architecture to support **high-throughput inference**, **eliminating I/O overhead** and ensuring system stability under heavy production loads.

GetAnime | Python, Langchain, Streamlit, Groq, ChromaDB, Docker, K8s, GCP | Repo

- Delivered sub-second query latency and **95% relevance accuracy**, architecting a Retrieval-Augmented Generation (**RAG**) pipeline with **LangChain**, **Groq LLM**, and **ChromaDB** to enable semantic search across **12,000+ anime entries**.
- **Maintained 99.5% system uptime** for the production environment, orchestrating containerized microservices via **Docker**, **Kubernetes**, and **GCP** while establishing **Grafana** dashboards for **real-time** health monitoring.
- **Optimized** retrieval context and search precision, engineering an automated ETL pipeline that utilizes Hugging Face Sentence Transformers to clean, merge, and generate vector embeddings for complex **multimodal** datasets.

RAGineer | Python, PostgreSQL, Ollama, Chainlit, ChromaDB, RAGAS | Repo

- **Boosted** SQL generation accuracy **from 42% to 84%** by architecting a **RAG** pipeline with **ChromaDB** retrieval of schema docs and query examples, enabling natural language **PostgreSQL** querying with **∼2.5s** average latency.
- Engineered specialized **code LLM** integration (**Qwen2.5-Coder**), achieving **0.91 retrieval precision** and **80% end-to-end accuracy** across complex multi-JOIN and subquery scenarios via semantic vector search.
- Hardened system for **production** with **SQL injection prevention**, rate limiting, and query complexity checks, validated by **64 unit & integration tests** at **97% coverage** using **RAGAS** evaluation framework.

Technical Skills

Languages: Python, SQL, Java, C
AI & Foundation Models: Generative AI, LLMs (GPT, BERT, Titan, Hugging Face), RAG Pipelines, LLM Fine-tuning
ML/DL: Computer Vision, Deep Learning, Machine Learning, Time Series, Statistics, NLP
Frameworks: PyTorch, TensorFlow, Keras, Transformers, LangChain, scikit-learn, Pandas, NumPy, Matplotlib
Cloud & MLOps: AWS (SageMaker), Azure, GCP, Kubernetes, CI/CD Pipelines, Git, GitHub
Data Engineering: PostgreSQL, MongoDB, Vector Databases (FAISS, Pinecone), Data Modelling
Software Development: Flask, Streamlit, RESTful API, FastAPI, OOP, DSA

Publications

Barbell Exercise Classification and Repetition Counting | ICDEC 2024 | Springer Nature | Link

- Engineered a robust machine learning system for barbell exercise classification and repetition counting using **MetaMotion sensor data**, achieving **over 90% accuracy** by developing comprehensive feature engineering and outlier detection pipelines to enable precise **human activity recognition**.

Honors & Awards

- **Dean's List** for Excellence in Academics **(highest GPA)** | Link
- **2 x Dean's List** for Excellence in Off-campus Achievements | Link