Dhruv Madhwal

linkedin | github | dhruvmadhwal@gmail.com | (623) 200-6557

EDUCATION

Arizona State University, Tempe, AZ **August 2024 - May 2026`**
Master of Science in Computer Science 4.11 GPA
Birla Institute of Technology and Science, Goa, India **August 2017-May 2022**
Master of Science in Physics 7.61 GPA
Bachelor of Engineering in Electronics and Instrumentation 7.61 GPA

TECHNICAL SKILLS

- **Programming Languages & frameworks:** Python, C/C++, MATLAB, Flask
- **Machine Learning:** PyTorch, TensorFlow, Keras, scikit-learn, Transformers, Hugging Face, OpenCV, pandas, NumPy
- **LLM/Agent & RAG Stack:** LangChain/LangGraph, AutoGen, ChromaDB, Pinecone, FAISS
- **Data Engineering & databases:** Kafka, Spark, Airflow, MySQL, Postgres, MongoDB
- **Cloud & Devops:** AWS, Docker, Git, CI/CD, MLflow

PROFESSIONAL EXPERIENCE

Data Scientist - Trukker Technologies **July 2022 - July 2024**

- Deployed a RAG-based customer support chatbot using AWS Bedrock, Pinecone, and AWS Lex, delivering instant query resolution (orders, billing, tickets) and reducing support volume by ~22%.
- Built dynamic pricing engine (XGBoost, LightGBM) trained on historical shipping data and external signals (fuel, demand etc.). Backtesting revealed ~12% improvement in quote acceptance.
- Automated email parsing pipeline leveraging GPT-3.5/LlamaIndex for structured data extraction from 4K+ monthly emails, achieving 96% extraction accuracy, reducing manual data entry by ~30%.
- Built a centralized Spark-based batch processing and ETL pipeline, consolidating data from Kafka, REST APIs, and databases into Redshift, reducing duplication by ~53%.
- Engineered a real-time invoice streaming system with Kafka & MongoDB Change Streams to sync financial data across multiple systems, lowering sync errors by ~12%.

Data Science Intern - Carelon Global Solutions **January 2022 - June 2022**

- Created an automated document classification pipeline by fine-tuning BioBERT on patient medical records, enabling the approval or rejection of prior authorization (PA) requests with 96% accuracy.
- Executed text extraction from medical documents using computer vision with OpenCV and PyTesseract, achieving 98% OCR accuracy.

AI/ML Engineering Intern - Samsung Research Institute **July 2021 - December 2021**

- Developed a TensorFlow Lite-based, context-aware recommendation system for Samsung app functionalities, leveraging user interactions and Bixby data to enhance personalization, achieving a 71% Hit Rate@5.
- Implemented MLflow & Airflow workflows, automating model versioning, deployment, and monitoring across multiple ML projects, enhancing efficiency and traceability.

RESEARCH EXPERIENCE

Graduate Researcher - CoRAL Research Lab at ASU **February 2025 - Present**

- **Multi-Hop Reasoning Agent:** Developing a LangGraph/AutoGen agent for open/closed-book QA that melds retrieval-augmented generation, question decomposition, and inference-time scaling. Simultaneously creating a novel in-house LLM-as-Judge evaluator to replace EM/ROUGE blind spots. Targeting new SOTA on FanOutQA, MuSiQue, and Frames.
- **Text-to-Infobox Sync:** Developed a multi-stage LLM pipeline that transforms free-text Wikipedia articles into up-to-date semi-structured infobox tables (extractive summarization → QA-SRL → knowledge-graph merge). Creating an evaluation suite for key-value accuracy and sync quality.

PROJECTS

Machine Unlearning in Small Language Models

- Implemented advanced machine unlearning techniques (gradient ascent, random labeling) on 3–4B parameter LLMs/SLMs, removing targeted knowledge while preserving overall performance (verified via BLEU/ROUGE-L).
- Evaluated four model variants (on Wikipedia_Person_Unlearn & TruthfulQA datasets) and identified Phi-3.5-mini-instruct as delivering the best trade-off between unlearning effectiveness and minimal quality loss.

Fine-Tuning Llama-2-7b-hf for Enhanced Mathematical Reasoning

- Fine-tuned Meta's Llama-2-7b-hf on GSM8K mathematical reasoning dataset via PEFT (LoRA, Adapters) and quantization methods, attaining a 31.4% exact-match accuracy, dramatically improving upon the ~5% baseline.

SQL-Native Fraud Detection Pipeline

- Embedded a TorchServe-hosted PyTorch CNN as PL/Python UDFs in PostgreSQL, fraud-detecting base-64 ID-card images in-place with 92 % accuracy, no ETL or micro-services, and triggered by a single SQL call.
- Enables fraud and non-ML analysts to surface high-risk IDs with a standard SQL query, significantly reducing operational complexity.