Aman Pratap Singh

linkedin.com/in/aman-pratap-singh54 github.com/amanpratapsingh54 aman16@umd.edu +1 (240) 960-6091

Education

University of Maryland — College Park, Maryland
Master of Science in Applied Machine Learning — Expected May 2026

National Institute of Technology — Hamirpur, India
Integrated Masters and Bachelors in Computer Science and Engineering — August 2017 – May 2022

Technical Skills

Languages: Python, Java, C++, Golang, R

Data Processing: Pandas, NumPy, SciPy, Matplotlib, Tableau

Libraries: TensorFlow, PyTorch, Keras, OpenCV, LangChain, OpenAI, DeepSpeed, scikit-learn

NLP & GenAI: GPT, LLaMA, Mixtral, NLTK, Reinforcement Learning, Feature Engineering

Cloud & Databases: AWS, Azure, GCP, SQL, MySQL, MongoDB, PostgreSQL, Redis

Backend & API Development: FastAPI, Flask, Django, Spring Boot, Datadog, Kafka, Docker

Experience

DeepEmergence — January 2023 – July 2024
Software Engineer — Bengaluru, Karnataka

- Fine tuned a domain specific **LLama-2** model on **60K** financial Q&A pairs and regulatory filings to power a recommendation engine. Boosted Top-3 retrieval accuracy by **37%** in offline tests and enhancing product relevance for wealth management users.
- Architected a low latency **RAG** system using **Hugging Face embeddings, FAISS IVF-PQ indexing**, and LangChain to ground **LLM** responses in **1,200+** financial PDFs; hallucination rates dropped **41%** and user trust scores rose 23% in internal testing.
- Integrated **Azure Document Intelligence** for OCR and key-value extraction, automating table, section, and layout parsing from unstructured financial docs; improved structuring accuracy by **33%** and speed up query pipelines by **19%**.
- Built an interactive portfolio analytics dashboard using **Plotly Dash** and **MongoDB**, backed by a **SARIMAX-XGBoost** ensemble model to predict equity trends. Achieved a Mean Absolute Percentage Error (MAPE) of **6.2%** across a diverse set of 50 stocks.
- Implemented unsupervised anomaly detection via **PCA** and **DBSCAN** clustering to flag irregular trading patterns, improving monthly rebalance signal precision by 17% over baseline heuristics.

Pristyn Care — June 2022 – December 2022
Backend Developer — Gurugram, Haryana

- Built internal automation tools (**Python, FastAPI**) to serve product metadata from MongoDB, and integrated **Google My Business** and WhatsApp APIs with custom **ETL pipelines**, reducing manual query load by **50%** and enabling real time data enrichment for segmentation tasks.
- Developed predictive systems for sentiment analysis, image triage, and no show forecasting using **RNNs, ResNet**, and time series models; boosted diagnostic throughput by **22%**, achieved **87%** accuracy, and optimized scheduling across 15+ specialties.

Projects and Research

Master Thesis - Speech Emotion Recognition | *Advisor Dr. Naveen Chauhan* — June 2020 – June 2022

- Built a multi speaker emotion recognition pipeline using **pyannote** audio for **speaker diarization** and VAD; extracted MFCC, Chroma, and Spectral Contrast features with augmentation (noise, pitch shift) to improve robustness on real world dialogue audio.
- Trained BiLSTM and CNN BiLSTM models with attention on IEMOCAP and CREMA-D datasets, using CTC loss for variable-length inputs; achieved **83%** accuracy across 6 emotions, outperforming SVM/RF by **10–12%** under 5-fold validation.

HybridMLComp: Optimizing ML Inference with Compiler Stacks — Jan 2025 – May 2025

- Built a benchmarking framework to evaluate **ML compilers (TorchInductor, ONNX Runtime, TensorRT)** across diverse models (**BERT, ResNet50**) and quantization methods (FP16/INT8), measuring latency, throughput, and memory usage.
- Optimized **hybrid compiler pipelines**, achieving up to **45%** higher throughput in large batch inference; results used to train a meta-compiler for intelligent backend selection.

Text-to-SQL Sequence-to-Sequence Modeling — January 2024 – May 2024

- Designed and implemented **Seq2Seq models** (basic, attention based, and attention with copy mechanism) to convert natural language questions into SQL queries, leveraging Spider and ViText2SQL datasets.
- Built custom training pipeline with TensorFlow 1.5 and **GloVe embeddings** on GPU; handled data preprocessing, vocabulary creation, and evaluated via execution accuracy.

S&P 500 Stock Forecasting — Sept 2024 – Dec 2024

- Developed hybrid ensemble combining **SARIMAX** time series models with LSTM neural networks using TensorFlow and statsmodels, achieving **15%** reduction in RMSE compared to individual models on 5 years of S&P 500 daily data.
- Assembled a robust feature pipeline incorporating historical price, volume, and technical indicators (RSI, MACD, SMA); leveraged **PCA** to condense 150+ features into 40 components, reducing training time by **20%** without sacrificing model accuracy.

PromtIQ: Generative AI-Enhanced Chat Application — January 2024 – Jun 2024

- Built a **Flask** based semantic QA application integrating **OpenAI APIs** to allow conversational querying over PDFs; parsed docs with **PyMuPDF**, embedded with OpenAI, and indexed using FAISS for fast, context aware LLM responses.