Manish Kumar

☎ +1 647-671-8526 | ✉ mak.manishkotra@gmail.com | in LinkedIn | GitHub

Experience

Classroom Ambassador - Data Science Sept 2024 – April 2025
Tech2U, University of Toronto Toronto, Canada

- Developed PoC of scalable multimodal **RAG pipeline** for automated information retrieval from large-scale text and visual datasets, integrating a ColPali-based vision-language neural retrieval module.
- **Achieved 10× faster retrieval** by optimizing ColPali via **4-bit QLoRA** fine-tuning, novel **32× patch-reducing pooling**, and high recall (∼**93% NDCG@10**) re-ranking.
- Developed **kernel-based embedding fusion** method extending late chunking, and robust **semantics-aware document parsing** pipeline using SmolDocling to improve LLM response grounding on long-form documents.
- Collaborated with SMEs and frontend engineers to translate domain-specific needs into chatbot response behaviors via **instruction-tuned prompt engineering** and **deployed FastAPI endpoints** for front-end chat UI integration.

Research Assistant Oct 2019 – Oct 2022
Artificial Intelligence and Robotics Lab, Indian Institute of Science Bangalore, India

- **Researched surrogate ML models** to approximate individual control surface moment contributions via multivariate nonlinear regression, **achieving <5% relative error** vs. wind tunnel data across wide flight envelopes.
- Developed **physics-informed ML models** integrating 6DOF equations to simulate aircraft dynamics, capturing complex nonlinear interactions while **ensuring sub-millisecond latency** for real-time closed-loop testing.
- Architected end-to-end simulation workflows with CI/CD pipelines for Software-in-the-Loop (SIL) and Hardware-in-the-Loop (HIL) environments, **reducing system validation and prototyping cycles by 40%**.
- Built **ELT pipelines** to clean large-scale noisy wind tunnel data; created **visualization dashboards** uncovering insights that informed design enhancements and improved aircraft roll performance by **8%**.
- Developed scalable **auto code generation pipelines** using MATLAB Coder, automating flight controller deployment and **reducing development time by 30%**.

Education

University of Toronto, Master of Engineering (Robotics) | GPA: 3.84/4 Jan 2023 – Jan 2025
Coursework: Machine Learning, Neural Networks and Deep Learning, Computer Vision, Reinforcement Learning, Big Data and Cloud-Based Data Analytics, Prompt Engineering

Skills

- **Programming Skills**: Python, SQL, MATLAB, OOPs
- **Technical Skills**: LLM Fine-Tuning, RAG, Generative AI, RL, 2D/3D Computer Vision, Distributed-training
- **Tools & Libraries**: PyTorch, Hugging Face, Transformers, Scikit-Learn, LangChain/LangGraph, PEFT, MLflow, Docker, Git/GitHub, FastAPI, Streamlit, Apache Spark
- **Cloud Platforms**: Azure (ML, Databricks, Data Factory), AWS (SageMaker, EC2), E2B Sandboxing

Projects

LLM-Powered Financial Sentiment Analysis & Trading System

- **Fine-tuned BERT-based model** and **embedding-based classifier** on the Financial PhraseBank dataset for sentiment analysis, compared its classification accuracy against the SOTA FinBERT model.
- Engineered ELT pipelines curating 3300+ days of financial dataset; performed **time series analysis** and **feature engineering** to extract time series and market sentiment based features, enabling robust ML model development.
- Designed an **AutoML** workflow integrating Ridge, Random Forest, and Gradient Boosting models; achieved a 40% improvement over baseline through automated model selection and hyperparameter tuning.

Cloud-Based Intrusion Detection System (IDS)

- Implemented ChiSqSelector for feature selection and trained multiple ML models using **distributed training** in **Apache Spark** to achieve a high-performance IDS (AUROC: 99.55%, AUPR: 96.24%) on KDD99 dataset.
- Developed an end-to-end MLOps workflow using **Azure Databricks**, **MLflow**, and **Docker** to support reproducible experiments, automated logging, and cross-environment portability.

Transformer-Based Text Summarization Microservice

- **Fine-tuned Google Pegasus** transformer model on Samsum dataset using **Hugging Face Trainer**, achieving 12% improvement in ROUGE-L and enhancing short-form text abstraction accuracy in domain-specific task.
- Built a full-stack MLOps pipeline encompassing data ingestion, model training, evaluation, and CI/CD integration with **Docker** and **GitHub Actions**, reducing deployment time by 50%.
- Deployed the summarization model as **RESTful microservice on Azure VMs** with **FastAPI backend** and **Streamlit UI**, enabling real-time user interaction and feedback-driven model refinement.

Self-Correcting Coding Agent

- Developed a modular self-correcting LLM coding agent using **LangGraph** for automated code generation and debugging, with multi-stage validation, runtime error handling, and iterative self-healing repair loops.
- Integrated **E2B sandboxing** to securely execute generated code in isolated environments, enabling automatic import resolution, runtime validation, and safe error recovery without human intervention.
- Combined **GPT-4o** for chain-of-thought (CoT) problem decomposition with a **RAG-powered CodeLLaMA** module, reducing function call overhead and improving code correctness in iterative problem-solving tasks.

Lightweight CNN for Image & Video Deblurring

- Designed a compact CNN-based Nested-ResNet (<9M params) using **PyTorch** based on extensive literature review; developed **layer-wise adversarial training**, outperforming DeblurGAN-V2 (60M) by 1.2 dB.
- Scaled training pipeline using **distributed training** (with mixed precision) across 2 heterogeneous GPUs, **enabling 200% larger batch size** and boosting overall training efficiency.
- Experimenting (on-going) blur-conditioned **Latent Diffusion** (multi-scale guidance) and **self-supervised perceptual similarity metrics** for out-of-domain generalization.

Event-Driven Document Indexing

- Designed and implemented an automated ETL pipeline using **Azure Data Factory** and **Azure Blob Storage** to process documents, generate vector embeddings with **Azure OpenAI**, and index them for downstream retrieval.
- Implemented an event-driven architecture using **Azure Functions** to monitor changes in **Blob Storage** and trigger incremental execution of the ETL pipeline, enabling near real-time document indexing.

Publications

- **Manish Kumar et al.**, "Analysis of an Axial Turbine using three different Vortex Laws" Proceedings of National Aerospace Propulsion Conference, Springer, Singapore, 2020.
- [Manuscript] **S. Sundaram, Manish Kumar et al.**, "Robust simultaneous stabilization based passive fault-tolerant controller for a fixed-wing aircraft".