ANSH KHANDELWAL

anshk@andrew.cmu.edu • 412-390-9196 • Personal Webpage • LinkedIn

EDUCATION

Carnegie Mellon University, School of Computer Science Pittsburgh, PA
Master of Science in Artificial Intelligence and Innovation, GPA: 4.0/4.0 May 2025

- Working on error-recovery of AI agents in collaboration with **Cohere** and Prof Graham Neubig
- Courses: Advanced NLP, Multimodal ML, AI Engineering, Generative AI, On-Device ML, Deep Learning Systems, Search Engines

International Institute of Information Technology, Hyderabad Hyderabad, India
Bachelor of Technology in Electronics and Communication, GPA: 8.8/10 May 2023

- Dean's Research Award (2022, 2023)

WORK EXPERIENCE

Kensho Technologies Cambridge, MA
Machine Learning Engineer Intern May 2024 – August 2024

- Evaluated LLMs and vision language models for financial document chart data extraction and question answering.
- Extracted structured data from line, scatter, and bar plots by customizing proprietary **Linked Objects Transformer** (a modified Detection Transformer-based image-to-graph model). (Applied for **Patent**)
- Productionized chart classification computer vision models (Vision Transformer, ResNet, and CNNs) with Airflow DAGs for automated training. Created a Python inference package and Grafana monitoring, attaining 7ms inference speed with 0.95 recall.

Siemens New Delhi, India
Machine Learning Engineer Intern July 2022 – August 2022

- Engineered a seq2seq LSTM-based encoder-decoder for time-series forecasting on Siemens' Symphony platform, achieving a 25% improvement in forecast precision. Optimized a C++ pipeline with python bindings, reducing model deployment time by 30%.

Samsung Bangalore, India
Software Engineer Intern May 2022 – June 2022

- Developed a C-based delay-detecting tool by logging camera system parameters, and implemented a Python graph-based analysis pipeline, reducing error-correction operational time by 80%.

RESEARCH EXPERIENCE

Language Technologies Institute, CMU Pittsburgh, PA
Research Assistant | Collaboration with **Cohere** with Prof Graham Neubig January 2025 – Present

- Creating synthetic data and setup **distributed fine-tuning** of Cohere's models, evaluating performance on WebArena to enhance LLM-driven web navigation and address error recovery challenges.
- Developing non-linear trajectory sampling to mitigate **exposure bias** and balance action-space for improving agent performance.

BNY Pittsburgh, PA
Machine Learning Engineer (Capstone Project) November 2024 – Present

- Building a **prompt optimization** suite to boost LLM performance in financial applications. Leading **prompt compression module**, by creating GPT-4 distilled datasets and fine-tuning models for token classification to achieve 20x compression ratio.

PUBLICATIONS AND PATENTS

- Quantitative Information Extraction from Figures in Documents, Applied for **Patent with Kensho**
- Improving IoT-based Smart Retrofit Model for Analog Water Meters using DL based Algorithm, **IEEE FiCloud 2022**
- Making Analog Water Meter Smart using ML and IoT-based Low-Cost Retrofitting, **IEEE FiCloud 2021**
- System and Method for Digitizing in an Analog Water Meter using Machine Learning, **Patent**: 202141021341

PROJECTS

MLOps for Movie Recommendation System January 2024 - April 2024

- Implemented a movie recommendation system serving **1 million** simulated customers using Content-based Filtering.
- Containerized with Docker to ensure scalability and 99% uptime through load balancing and automated retraining with Jenkins. Utilized MLflow for model versioning and set up real-time monitoring and metrics tracking with Prometheus and Grafana.

End-To-End Large NLP System Building with RAG January 2024 - April 2024

- Designed an NLP system for QA with advanced retrieval-augmented generation (RAG), creating a baseline model using FlanT5 with LangChain and a FAISS database, resulting in an 8s inference speed on a local device.
- Built a Chroma database with HuggingFace UAE-Large embeddings for document retrieval and utilized a quantized Llama-2-7b model with Ollama, improving ROUGE scores by 60% over baseline.

Multimodal Interaction in Memes January 2024 - April 2024

- Designed a meme caption generation pipeline by fine-tuning InstructBLIP, MiniGPT4, and Facebook MMBT as baselines.
- Enhanced caption generation by using CLIP embeddings-based retrieval technique and integrating external knowledge, resulting in a 15% increase in ROUGE scores through both intrinsic and extrinsic evaluation.

SKILLS

Programming: Python, C, C++, Java, JavaScript, Bash, PyTorch, TensorFlow, Pandas, Numpy, VLLM, LlamaFactory
Development: AWS Sagemaker, Kafka, Jenkins, Grafana, Docker, Kubernetes, GCP, MLFlow, Apache Airflow, DVC