ANVESHA KATARIYAR

Email: aakatari@alumni.cmu.edu | Contact: 412-456-7890 | Website: https://aakatari.github.io/

EDUCATION

Carnegie Mellon University — Pittsburgh, PA
Master of Science, Information Systems — AI Concentration | ***GPA**: 3.8/4.00* — *Aug. 2023 – Dec. 2024*

- **Courses:** Generative AI, Advanced NLP, Multimodal ML, Search Engines, Distributed Systems, DBMS, Statistics
- **Teaching Assistant:** Machine Learning, Applications of NL(X) & LLM, Decision Analytics for Business

Cummins College of Engineering, Savitribai Phule Pune University — Pune, IN
Bachelor of Technology, Electronics & Telecommunication | ***GPA**: 3.72/4.00* — *Aug. 2016 – Aug. 2020*

- **Courses:** Data Structures, OOPS, Big Data & Analytics, Operating Systems, Computer Networks, Signal Processing

TECHNICAL SKILLS

Programming Languages C++, Python, C, Java, C#, R, SQL, MATLAB, Shell Scripting, Go, Scala, JavaScript
Libraries & Frameworks PyTorch, Tensorflow, Keras, NLTK, OpenAI, Transformers, LangChain, Hive, AWS, GCP, Azure, Spark, Hadoop, Kafka, PostgreSQL, Kubernetes, Docker, Tableau, A/B Testing

EXPERIENCE

Tata Consultancy Services, Road Intel — Pittsburgh, PA
Machine Learning Engineer — *Aug 2024 – Dec 2024*

- Developed an AI system for Level 3+ autonomous vehicles, integrating YOLOv8 for object detection (potholes, pedestrians, vehicles), LSTM for trajectory prediction, and custom algorithms for pothole severity estimation.
- Engineered a cost-efficient, real-time camera-based ML module with 90% reduced computational overhead, incorporating CV algorithms for pothole detection using monocular depth estimation, achieving 85.47% accuracy in size prediction.

Carnegie Mellon University, OpenHands — Pittsburgh, PA
Artificial Intelligence Researcher — *Jun 2024 – Dec 2024*

- Designed a modular multi-agent framework for code generation, enabling LLM agents to collaboratively retrieve, assemble, and adapt task-specific workflows dynamically based on user prompts.

Carnegie Mellon University — Pittsburgh, PA
Artificial Intelligence Research Intern — *Jun 2024 – Sep 2024*

- Fine-tuned LLaMA-3 and Mistral models using a RAG pipeline on 1,000+ job postings across 10 domains, achieving a 30% improvement in content ranking accuracy and generating personalized CVs and cover letters with high ROUGE scores.
- Built a multi-turn, context-aware chatbot using LoRA and QLoRA-tuned LLaMA-7B integrated with LangChain, enabling real-time personalized conversations with dynamic memory and prompt adaptation.

Western Digital Corporation — Bengaluru, IN
Senior Data Scientist — *Aug 2020 – Aug 2023*

- Built time-series regression models (XGBoost, linear) to forecast wear trends from terabyte-scale SSD telemetry, enabling short-stroke optimization, reducing validation cycles by 30%, and contributing to $13M in accelerated product revenue.
- Applied unsupervised learning techniques (K-Means clustering, DBSCAN) on endurance and failure logs to detect latent failure modes, enabling early risk stratification and refinement of reliability baselines across engineering teams.
- Developed scalable ETL workflows using Python and SQL to preprocess and validate 4M+ telemetry records, improving data quality by 11% and accelerating downstream modeling and experimentation cycles.
- Built interactive visualization tools using Dash/Plotly and JMP to analyze SSD wear and degradation patterns, reducing root cause analysis time by 8 days and improving visibility for cross-functional stakeholders.

PROJECTS

Generative AI: Go Small or Go Home | *CMU* — Aug 2024 – Dec 2024

- Explored model compression techniques quantization, pruning, and knowledge distillation to optimize generative models for efficient deployment on edge devices, reducing memory footprint and inference latency across text and image generation.

Twitter Recommendations System with Apache Spark | *CMU* — Jan 2024 – Apr 2024

- Built a scalable Twitter topic recommendation system, ranking topics for 10M users based on PageRank scores in a large social network graph, utilizing Azure Databricks, Scala, and Spark SQL.

Movie Recommendations System with Apache Kafka | *CMU* — Jan 2024 – Apr 2024

- Kafka-based pipeline to process 1M+ daily user-movie interactions, enabling real-time ML via feature engineering.
- Built CI/CD pipelines for automated retraining and deployment, improving model tracking and uptime by 15%. Added content-based filtering using KNN on metadata embeddings, validated via A/B testing, boosting user interaction by 3.5×.

Think Fusion: Augmented Outside Knowledge Visual QA | *CMU* Aug 2024 – Dec 2024

- Developed ThinkFusion, a VQA model that achieved 83.85% accuracy by integrating code generation, free-text rationale, and custom data augmentation. Used ViperGPT to inject commonsense context and free-text rationales, improving model personalization, ranking quality, and robustness in complex multimodal inputs.

Multilingual Multicultural Figurative QA | *CMU* Jan 2024 – Apr 2024

- Enhanced encoder-decoder models with LoRA fine-tuning, contrastive loss training, and data augmentation. Improved context understanding, achieving an 11% boost in cross-cultural communication accuracy.

Search Optimization with BERT | *CMU* Jan 2024 – Apr 2024

- Constructed a Search Engine in Python utilizing BERT. The re-ranker was indexed on Lucene and powered by a BM25 ranker on QueryEval-based data, delivering a mean average precision (MAP) of 0.577 and a Recall of 93%.