Ankita Tripathi

+1 (929) 690-3533
ankitatripathi95@gmail.com
GitHub | LinkedIn

EDUCATION

Master of Science in Computer Science (AI concentration) September 2024 – May 2026 (Expected)
***New York University**, New York, USA*

Bachelor of Science in Computer Science GPA: 3.7 August 2017 – May 2021
***University of Central Florida**, Orlando, USA*
- Minors in Intelligent Robotic Systems and Cognitive Sciences

EXPERIENCE

Software Development Intern January 2023 – March 2023
FULL Creative, Chennai, India
- Implemented a login functionality proof-of-concept with 15% reduction in latency through optimized Java Servlet implementation and gained working knowledge of the Spring framework and MVC architecture.

Undergraduate Research Assistant August 2020 – May 2021
Department of Biology, University of Central Florida, Orlando, USA
- Developed a Keras-based Convolutional Neural Network (CNN) for Amblyomma americanum (lone star tick) image classification, achieving 85% accuracy.
- Collected and preprocessed image dataset, implementing data augmentation and image segmentation techniques to enhance model training and achieve optimal classification accuracy.

PROJECTS

LLaMA RAG System: Retrieval-Augmented QA February 2025 – April 2025
Big Data and Machine Learning Systems, New York University
- Implemented an end-to-end RAG pipeline around a fine-tuned LLaMA-3B, including document ingest + chunking (300/50), dual embedding backends (all-MiniLM-L6-v2, bge-large-en), and five FAISS index types (Flat, IVF, PQ, HNSW, IVF-PQ) to ground answers in retrieved context.
- Benchmarked 5×2 retrieval configs on a climate QA set (10 queries); retrieval latency ranged 8–124 ms by index/embedding while generation dominated >99% of total time (~20.5–20.9 s vs baseline 20.49 s), showing RAG adds minimal end-to-end overhead.
- Introduced a token-efficiency metric (time/token): RAG responses were +1–9% slower due to larger context but delivered more specific, grounded outputs; recommended bge-large-en + HNSW as the quality–speed sweet spot and documented trade-offs.

Distributed LLaMA-3B Fine-Tuning February 2025 – April 2025
Big Data and Machine Learning Systems, New York University
- Fine-tuned the 3 B-parameter LLaMA 3.2 model on a 1 M-token climate-documents corpus using LoRA + NF4 4-bit QLoRA, trimming trainable weights to 0.06 **%** and fitting an effective batch of 64 on a single A100 GPU.
- Extended training to two GPUs with DeepSpeed data, tensor, and pipeline parallelism, cutting epoch time 62 % (2658 s → 1002 s) while keeping perplexity < 12.
- Built reproducible torchrun + Slurm workflow and built an evaluation suite that logs loss, perplexity, and GPU utilization; best data-parallel run achieved eval loss 2.16 / ppl 8.70

Replicated Concurrency Control and Recovery System September 2024 – December 2024
Advanced Database Systems, New York University
- Built a distributed database with serializable snapshot isolation (SSI) and Available Copies replication in Java.
- Designed a transaction manager ensuring consistent distributed operations and failure recovery.
- Developed dependency-graph-based validation with cycle detection to prevent anomalies.
- Implemented multi-version concurrency control (MVCC) with snapshot isolation.

SKILLS

Python, Java, C++, JavaScript | PyTorch, TensorFlow, NumPy, Pandas | Spring, Docker, Git