Venkata Harsha Pedada

Austin, Texas • Portfolio • (312) 459-8539
venkataharshapedada9@gmail.com • linkedin.com/in/venkata-harsha • github.com/Harshaiiitk

Objective: Recent graduate specializing in machine learning systems. Experienced in optimizing Large Language Model (LLM) inference for efficiency and scalability. Proficient in distributed LLM fine-tuning and inference techniques, utilizing custom CUDA kernels.

EDUCATION

Illinois institute of technology | Chicago, IL — *May 2024*
Master's in Computer Science | *CGPA*: 3.8/4

Indian Institute of Information Technology | Kottayam, India — *May 2022*
B.Tech in Computer Science and Engineering

TECHNICAL SKILLS

Languages: C, C++, CUDA, Go, Python, SQL, YAML, Bash, Makefile.
Frameworks: Pytorch, Ludwig, Huggingface, vLLM, Transformers, Sagemaker, Quantization.
Technologies: Linux, Kubernetes, Docker, Git, AWS, GCP, CI/CD, GPU Architecture, Parallel Computing.
Coursework: Multicore Programming, Deep Learning for NLP, Machine Learning. Data Structure and Algorithms

PROJECTS

Concurrent Data Structures in C++ — Mar 2024

- Developed high-performance concurrent data structures using C++ and shared locks, demonstrating expertise in multi-threading and synchronization primitives.
- Optimized for read-heavy workloads, achieving an ***800ms performance improvement*** for 1 million random reads over sequential implementations.

Image Super-Resolution — Jan 2024

- Engineered a ***4x SRGAN model*** in Python with Keras, integrating convolutional and residual blocks to enhance image resolution, achieving 87% validation accuracy.
- Optimized data pipeline for high-dimensional images using the MIRFLICKR-25000 dataset, reducing the preprocessing time by 20% with advanced augmentation for 128x128 inputs.
- ***Boosted model robustness and perceptual clarity by 30%*** through VGG19-based content loss and GAN architecture, achieving 72% testing accuracy and surpassing baseline CNN methods for super-resolution tasks.

Twitter Hate Speech Recognition — Dec 2023

- Developed real-time Twitter hate speech detection models in Python using scikit-learn, NLTK, and TensorFlow, with data extraction via Twitter API.
- Enhanced model accuracy by engineering features like n-grams, TF-IDF vectors, and sentiment scores to capture hate speech patterns effectively.
- ***Achieved 87% accuracy and 0.85 F1 score*** through optimized Logistic Regression with comprehensive text preprocessing, including tokenization and lemmatization, and improved model performance by 20% via hyperparameter tuning.

PROFESSIONAL EXPERIENCE

Principal Software Developer | Big Brothers Big Sisters at Ventura County — May 2024 - Present

- Implemented data processing workflows for volunteer reports, ***reducing retrieval latency by 25%***, enabling quick access to information on logged hours, and *improving admin efficiency*.
- ***Boosted reporting efficiency by 50%*** with automated workflows using Google Sheets and Drive APIs, and created visual data representations to track engagement trends over time.
- Led secure login and check-in system development with Flutter, Node.js, and MySQL, ***enhancing usability by 30%*** through iterative testing and performance optimization.
- Created search and filter features for reports, ***accelerating report generation by 40%*** and allowing admins to track metrics like volunteer participation and event attendance.
- ***Saved 8 hours weekly*** by automating email reminders and report deliveries, significantly reducing manual workload for staff and allowing focus on higher-priority tasks.

Teaching Assistant (Database organization) | Illinois Institute of Technology — Aug 2023 - May 2024

- Orchestrated and executed diverse teaching methods and assessments including interactive tutorials; provided detailed project feedback and strategic guidance to over 500 students, resulting in a ***15% increase in overall student performanc***e metrics.