Rahul Reddy Chappidi Venkata

2485613908 • rahulreddycv@gmail.com • linkedin.com/in/rahulreddycv • https://rahulcvr.github.io

ML Engineer with professional experience, specializing in building AI solutions for business problems, optimizing LLMs and building advanced agentic RAG systems. Proven ability to bring AI models from the lab to tangible products that reach millions of users.

WORK EXPERIENCE

Machine Learning Engineer

University of Illinois at Urbana-Champaign, (Cline Center for Advanced Social Research) **06/2024 - Present**

- Optimized and fine-tuned transformer models through quantization and pruning techniques with TensorRT, achieving a 30% decrease in size, 20% decrease in inference latency with only 3% drop in performance on the downstream task.
- Implemented **A/B testing with DiD analysis** on Prodege annotation workflows, measuring the impact of LLM assistance across 160 annotators and demonstrating a **28%** increase in throughput and a **39%** reduction in error rates.
- Optimized training of BERT and open-source LLMs on H100 GPUs with DeepSpeed and LoRA, boosting accuracy 15% and cutting training time 60%.
- Implemented an automated evaluation pipeline in MLFlow that ingests human-annotated examples from Prodege as they arrive, computes metrics and publishes performance dashboards, reducing manual evaluation effort by 80%.

Machine Learning Engineer

University of Illinois at Urbana-Champaign **03/2024 - 05/2024**

- Developed a RAG-powered chatbot on employee handbook using LangChain, FAISS vector search, and Olama models for context-aware policy retrieval.
- Optimized LLaMA models for downstream tasks through quantization and structured pruning, achieving a **40% decrease in size**, without significant drop in performance.

Machine Learning Engineer

Vedantu **07/2022 - 10/2022**

- Fine-tuned and deployed a DistilBERT model using **PyTorch** and **FastAPI** to classify and route student queries to tutors in real time, **reducing response time by 65%** and **increasing chat engagement by 3.5×**.
- Designed and executed **A/B testing** on **1M** users, across landing pages, and homepage, driving a **4% increase in conversion rates** over the next 3 months by improving UI/UX elements, simplifying navigation and reducing friction in sign-up funnels**.**
- Automated **Tableau** dashboards integrating **weekly data from 50M+ events** stored in **BigQuery**, using Airflow-orchestrated **SQL** and **Python ELT** pipelines to deliver KPI visibility to product and engineering teams, enabling faster decisions for feature rollouts.

SKILLS

Airflow, AWS, CI/CD, CUDA, DeepSpeed, Docker, Git, Hugging Face, JAX, Jenkins, Jupyter, Kubernetes, Langchain, Matplotlib, MCP server, Optuna, Python, PyTorch, RAGs, TensorFlow, TensorRT

EDUCATION

Masters of Science in Information Management

University of Illinois, Urbana-Champaign • GPA: 4.0/4.0 05/2025

Bachelors of Technology in Computer Science and Engineering

Jawaharlal Nehru Technological University Hyderabad • GPA: 3.77/4.0 05/2023

Relevant Coursework: Text Mining, Advanced Statistics, Applied Machine Learning, Generative AI, Natural Language Processing

PROJECTS

RAG Based Knowledge Management

- Open-sourced a Python API for RAG-based chatbots over personal notes, leveraging Chroma DB vector store and transformer embeddings to deliver >90% answer accuracy with inline source citations.
- Designed a lightweight REST API with Streamlit UI for natural-language querying of knowledge bases, achieving <150 ms latency.

UIUC Gym Occupancy Detection

- Built a density-sampling gym occupancy estimator on Raspberry Pi: defined a homography-corrected ROI, sampled 5 % patches, counted heads via simple blob analysis, and extrapolated to full-floor occupancy in <10 ms/frame.
- Deployed the edge pipeline on Raspberry Pi with OpenCV & NumPy, powering a real-time gym capacity dashboard at sub-50 ms end-to-end latency and ±15 % accuracy—no heavy DNN required.