Bhargav Kantheti

332-699-0585 | Linkedin | www.bk.bio | Github | bk2899@columbia.edu

Skills

Python, PyTorch, Ollama, Kubernetes, Docker, AWS, NextJS, ReactJS, SQL, Redis, MongoDB, Dask, PySpark, Git, Linux.

Experience

NYC Administration for Children's Services New York, NY
Research Scientist (Consultant) *Sep 2024 – Present*

- Developed a retrieval-augmented generation pipeline for Azure PaaS to support first responders with real-time insights, reducing hallucination rates by 45% through rigorous citation enforcement, optimized document indexing using local VectorDB, and topic modeling.
- Developed and deployed frontend, telemetry, Express API, and MongoDB microservices on a short timeline for the RecSys while ensuring high container CPU efficiency. Analyzed Grafana logs to automate container scaling, shrinking infrastructure costs by 10%.
- Devised NextJS/Redis/K8s apps to ensure fallback on model drift from feedback polls, preventing inaccurate outputs and enabling automated rollback of affected predictions in 70% of edge cases.

Columbia Engineering (Earth Institute) New York, NY
Capstone Research Assistant *Sep 2024 – Dec 2024*

- Researched transformer embedding techniques applied to over 70 years of climate data for latent anomaly detection, laying the foundation for novel environmental ML applications.
- Implemented an autoencoder-based data reversible compression model that reduced storage requirements by 60%.

NYC Administration for Children's Services New York, NY
Summer Graduate Intern *Jun 2024 – Aug 2024*

- Engineered a timeseries forecasting model that analyzed historical patterns to predict staffing needs with 80% accuracy measured over 3 quarters, identifying potential resource gaps before they affected service delivery.
- Designed interactive telemetry dashboards for the commissioner of ACS that translated ML applications in child welfare to the California leadership summit.

Columbia Medical Center (Pathology Lab) New York, NY
Research Assistant *Jun 2024 – Aug 2024*

- Analyzed system logs to identify security and performance vulnerabilities, implemented an optimized ETL pipeline that improved system integrity while increasing processing capacity by 50%.
- Engineered a cross-platform, user-centric ETL workflow designer that captures and operationalizes spur-of-the-moment research ideas, enabling agile innovation.

Feynn Labs Guwahati, IN
Machine Learning Intern *May 2022 – Jul 2022*

- Conducted market segmentation analysis for local retail stores by leveraging transactions and inventory logs. Identified trends to reduce overstocking and enhanced targeted customer engagement by 20% for small businesses.
- Led a predictive analysis project to forecast market viability for an EV company in India. Integrated battery supply chain dynamics with stock volatility analysis, yielding projections awarded top honors among six competing teams.

Education

MS in Data Science New York, NY
Columbia University *Aug 2023 – Dec 2024*

BTech in Computer Science and Engineering Visakhapatnam, IN
Gandhi Institute of Technology and Management *Jul 2019 – Apr 2023*

Projects

5ki: Built a unique RAG technique for quantized LLMs and digested Wikipedia to demonstrate viability. [blog]

intxr.net: Developed an Actor-Critic Reinforcement LLM model that creates a repository of free browser software. [blog]

relatable: Created a geospatial storytelling paradigm to bridge data insights with powerful impact. [blog]