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*

- Engineered an RAG pipeline for Azure PaaS by integrating topic modeling with a local VectorDB. Solution reduced inference latency by 70% and hallucination rates by 45%, delivering real-time insights for first responders.
- Architected a robust hierarchical recommendation system and LLM-inspired contextual embeddings to balance imbalanced datasets, enhancing rehabilitation service matching accuracies by 35–55% for minority classes.
- Designed and deployed a full-stack MLops ecosystem—including real-time telemetry, Express API, and MongoDB microservices—with auto-scaling via Grafana alerts, reducing infrastructure costs by 10%.

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.
- Developed 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*

- Developed 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.
- Conducted a comprehensive cost-benefit analysis on a $1M+ legacy service rebuild and proposed an ML-powered solution that reduced costs by 30% while doubling operational efficiency.

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.

GITAM (CS Lab) Visakhapatnam, IN

Research Assistant *Aug 2022 – Apr 2023*

- Constructed a fine-tuned CNN architecture for Telugu language OCR, achieving 98% accuracy, initiating new research avenues for the language.
- Authored an intuitive Python library combining IBM Qiskit and Google CirQ to aid Quantum Computing adoption across campus and empowering research initiatives.

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]