Kartik Choudhary

kartikchoudh@umass.edu | linkedin.com/in/kartik727 | Amherst, MA

EDUCATION

University of Massachusetts, Amherst, MA **May 2024**
MS in Computer Science GPA: 4.00/4

Indian Institute of Technology Delhi, India **Jun 2019**
B.Tech in Electrical Engineering GPA: 3.30/4

SKILLS

Tools and Languages : Python, C++, SQL, R, Java, Bash, LaTeX, LangChain, XGBoost, CUDA, Jupyter, Git, Linux
Big Data and Cloud : Spark, Hadoop, Hive, PySpark, Tableau, Streamlit, Docker, Kubernetes, AWS, GCP, Azure
Data Science and ML : PyTorch, TensorFlow, ONNX, HuggingFace, Keras, MLFlow, InterpretML, MongoDB, Postgres

PROFESSIONAL EXPERIENCE

Machine Learning Engineer, Reliance Jio Infocomm Ltd. **Jul 2021 - Jul 2022**
Led the initiative to develop cost-effective solutions for improving the 4G experience for customers and providing network insights.
- Coordinated a team of 6 for optimizing network coverage using *RSM* leading to 1.2M fewer daily calls with network issues.
- Attained a *15%* improvement in network throughput predictions by implementing a *Mixture of Experts* model with PyTorch.
- Overhauled the call drop RCA pipeline by utilizing *SHAP scores* resulting in 18% fewer false positives in downstream tasks.

Data Scientist, Reliance Jio Infocomm Ltd. **Jul 2019 - Jun 2021**
Used machine learning and Bayesian statistics to gain actionable insights into customer behavior and refine retention strategies.
- Deployed a distributed model with *Apache Spark* and *MS Azure* to perform churn prediction for *400M+* daily customers.
- Curated *over 1100* features for churn prediction to improve customer retention resulting in *$35M/mo* in additional revenue.
- Collaborated with vendors to launch *A/B testing* of ad campaigns on *geographic cohorts* and created a Tableau dashboard.
- Reduced data pre-processing time by *70%* during ETL by implementing a *multiprocessing pipeline* for feature extraction.

RESEARCH AND INTERNSHIPS

Machine Learning Intern, Meta **Feb 2024 - Present**
- Developed evaluation metrics for judge LLMs and studied the effects of their size and architecture on their performance.
- Designed and conducted experiments to assess the impact of instruction alignment and adversarial inputs on LLMs.

Data Science Intern, Microsoft **Jan 2024 - Feb 2024**
- Added generative model support to Responsible AI Toolbox with evaluation metrics and hierarchical importance measures.
- Created a benchmarking framework for prompt templates and did user studies to enhance the Azure AI Studio and Copilot.

Graduate Student Researcher, University of Massachusetts Amherst **Dec 2022 - Jan 2024**
- Developed off-policy algorithms to learn personalized sepsis treatment policies from historical data of ICU patients.
- Evaluated comparative reasoning capabilities of vision-enabled LLMs like GPT-4V by creating a novel dataset. [Best Project]

PUBLICATIONS

Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges, Under review, NeurIPS **2024**
A study of the properties of LLMs as judges in comparison to humans and automated evaluation methods. This work provides a comprehensive analysis of popular LLMs such as *GPT-4* and *Llama 3*, and potential pitfalls in their use as judges. [First author]

ICU-Sepsis: A Benchmark MDP Built from Real Medical Data, Reinforcement Learning Conference **2024**
Modeled sepsis management as an indefinite-horizon Markov decision process and utilized the EHRs of over 17k ICU patients to estimate the parameters of the MDP, which can be used to accelerate applied RL research in sepsis treatment. [First author]

PROJECTS

TaskLLM: Multi-Agent Planning and Execution **2024**
A multi-agent system with a *task-oriented language model* that can generate and execute plans in a simulated environment.

AI Writing Copilot **2024**
A writing assistant app built using *LangChain* and *FastAPI* with customizable levels of feedback and suggestions using *RAG*.

Cloud-based distributed stock exchange **2023**
Distributed stock exchange with a *microservice* architecture and REST interface hosted on *AWS*. Utilizes LRU cache for improved performance and replication for fault tolerance. Uses *Docker* for containerization and *Kubernetes* for container orchestration.

Darry the customer service bot **2023**
A *prompt-engineered* e-commerce customer service bot built using *ChatGPT* that can interact with the company's database.

HONORS AND ACTIVITIES

- Graduate Teaching Assistant, CS 687 and CS 550, UMass Amherst 2023-24
- Highest annual performance feedback rating (A*) for 3 years in a row, Reliance Jio 2020-22