Ritika Kumari

Machine Learning Engineer

✉ ritikakumari1302@gmail.com ☎ +1 404 4493530 Atlanta, USA medium linkedin riti1302

Professional Experience

Machine Learning Engineer III, *Walmart* 05/2024 – 07/2025 Bangalore, India

- Enhanced the efficiency of existing machine learning pipelines by **identifying bottlenecks** and implementing optimizations.
- Led the deployment of **large scale** machine learning models into production environments, ensuring they were **scalable, robust, and secure**.
- Developed the **forecast pipeline of a critical model** for optimum floor price **prediction in digital ads auction**.
- Built a central monitoring platform for data validation, **drift detection** and Root Cause Analysis (RCA), significantly reducing model error detection and response time.

Machine Learning Engineer II, *Glance* 01/2023 – 04/2024 Bangalore, India

- Full Stack MLE solving for content personalisation of Glance feed for **200 million** daily engaged users (DEU).
- Developed and implemented machine learning models such as logistic regression, random forest, xgboost, collaborative filtering, clustering etc for user profiling, content recommendation, and user behavior prediction resulting in a remarkable **23% increase in the time spent** on Glance platform.
- Write and manage ML services that can handle **80k QPS** with p99 **latency of 250ms** and create production jobs using KubeFlow, leveraging the modern day MLOPS stacks around CI/CD, featurestores and monitoring.
- Performed comprehensive Exploratory Data Analysis (EDA) to derive valuable insights from substantial user engagement data and metadata, employing **inferential statistics** and visualization techniques.
- Designed and implemented scalable data pipelines and processing systems to handle billion scale user data and deliver real-time personalized content.
- Collaborated with cross-functional teams to define and prioritize personalization requirements and design data-driven solutions.
- Technologies: Python, PySpark, Kafka, Hive, Hudi, GCP, Azure, GoLang, Kubernetes, Big Query

Data Scientist, *Glance* 07/2021 – 12/2022 Bangalore, India

- Implemented an end-to-end user level recommendation model for highly interactive users surpassing all other models and serving **70%** of the total user base (**140 million users**).
- Developed and deployed Hybrid Collaborative Filtering based approach for cold users, increasing the **monetisation metric (dollar value) by 20%.**
- Designed and developed a Clustering based Look Alike model, learning to extrapolate the behaviour of highly interactive users on **cold and sparse users**, increasing the **engagement metrics by 15%.**
- Implemented unit, integration and regression tests for Spark based ETL jobs

Data Science Intern, *Fidelity Investments (AI CoE)* 01/2021 – 06/2021 Bangalore, India

- Built a Knowledge Graph for closed domain financial data.
- Explored and implemented different techniques for entity and relation extraction from Natural Language Text.
- Developed a highly scalable novel search algorithm by leveraging the power of Knowledge Graph, **bringing down the search time per query by 25%**
- Technologies: Natural Language Processing, Neo4j, Knowledge Graph, Graph Embeddings

Data Science Intern, *Delhivery* 10/2020 – 12/2020 Gurgaon, India

- Improved the accuracy of the auto-verification pipeline of delivery addresses using a rule-based approach.
- Implemented a word-embedding model with sub-word information for spelling correction of addresses.
- Developed a module for bigram validation using Point-wise Mutual Information (PMI) and elastic search

Projects

Flexible LLM Inference with Multi Model Prefill and Decode , *LLM Inference (Research)* 08/2025 – Present

- Implemented a stitched LLM architecture using separate models for prefill and decode phases, improving both TTFT and TBT while maintaining model accuracy.
- Achieved **5% decrease** in **latency** than the baseline model (using same bigger model in both the phases)
- Tested out generalization capability for the architecture on different tasks (q/a, summarization, code completion, math problems)

Physics-Informed Neural Network for Aerodynamic Analysis, *Scientific ML* 08/2025 – 12/2025

- Built a PyTorch-based point-wise neural network with physics-informed loss constraints that maps spatial coordinates and design parameters to aerodynamic flow fields.
- Transformed a data-limited problem (500 configurations) into 35M training samples through pointwise prediction, achieving fast surrogate modeling for CFD simulations with **$R^2 > 0.95$**.

Education

Georgia Institute of Technology, *Masters of Science in Computer Science* 08/2024 – 04/2026
SGPA 4.0/4.0 Atlanta, USA

Indian Institute of Information Technology Kalyani, 2017 – 2021
Bachelor of Technology in Computer Science and Engineering India
CGPA 8.8/10.0

Skills

Machine Learning | Deep Learning | Recommendation System | GPU Programming | Natural Language Processing | Computer Vision | Knowledge Graph | Data Analysis | Data Visualization | Statistical Modeling | Data Engineering | Cloud Computing | Experimental Design | Apache Spark | Kafka | BigQuery | Neo4j | Jenkins | Airflow | LLM Inference

Awards

Winner, *Smart India Hackathon* 05/03/2019
Developed an end-to-end pipeline to search relevant books/journals from the database given 4-5 keywords

Real Star, *Glance Data Science Team* 13/10/2022

- Owned the most critical dense user modelling
- Improved 15% on engagement metrics
- High ownership