Rishabh Thapliyal

+1 858-281-1321 San Diego, CA rthapliyal@ucsd.edu LinkedIn Github

Education

University of California San Diego | *GPA: 4.0/4.0* Sep 2024 - Present
Master of Science in Machine Learning & Data Science *expected graduation Dec '2025*
Courses: Probability & Statistics for Data Science, Programming for Data Analysis, Statistical Learning I

Indian Institute of Technology (IIT) Bombay | *GPA: 8.4/10* Jul 2018 - May 2022
Bachelor of Technology in Chemical Engineering | Minor in Artificial Intelligence & Data Science *Mumbai, India*
Courses: Machine Learning, Deep Learning, Machine Learning for Remote Sensing, Mathematical Optimization

Technical Skills

Languages: Python, PySpark, C++, R, SQL
Developer Tools: GitHub, PyCharm, Docker, Streamlit, Kubernetes, Airflow, Postman, Confluence
Cloud & Big Data: GCP Vertex AI, Azure AI Studio, Spark, Hadoop, Cassandra, Redis, Google BigQuery
Libraries, ML Frameworks, & GenAI tools: Pandas, NumPy, seaborn, PyTorch, sklearn, Tensorflow, Transformers, HuggingFace, Dask, CUDA, LangChain, LlamaIndex, LLM Finetuning, LLM Evaluation, Prompt Engineering, RAGs, LLM Agents, PEFT, LoRA

Work Experience

Walmart Global Tech Jun 2022 - Sep 2024
Data Scientist III | Intl Catalog Team and Intl Global Sourcing Team *Bengaluru, India*

- Innovated a novel architecture by integrating **semantic router chains**, multiple LLM sub-chains, & **multi-prompting** techniques in SQL chains library of LangChain to build interactive **Text-to-SQL** queries chatbot with **90%** accuracy
- Designed and implemented an LLM-powered **retrieval augmented generative** question answering service for Walmart associates using GPT-4, LlamaIndex, LangChain, HuggingFace embeddings, Redis, and FAISS vector database
- Build a Global Item Mapper solution to map product reviews by using product title, descriptions, image features (HSV, SSIM, SIFT, & ORB), and attributes (brand, size, color, etc.) for **200M+** items across US, Mexico, and Canada. Scaled this pipeline across multiple **GPUs** (4 T4s) using **Dask** to reduce the computation time by **60%**
- Improved the catalog hierarchy for the Walmart's Mexico market by predicting product types from product titles for **10M+** un-navigable items using a max voting **ensemble model** of XGBoost, MPNet, and GPT-3.5
- Created a content quality scoring pipeline to score titles, images, and attributes for **50M+** products in the Walmart's Mexico catalog. Deployed this as an API on **Google Cloud Platform** and scaled it to handle **100+** requests/second
- Identified trends, seasonality, & cyclicity in the time series data of **20+** KPIs for **8k+** suppliers. Grouped these time series using **Dynamic Time Warping** and created forecasting models for each of them using ARIMAX, Prophet, LightGBM, and **hybrid forecaster model** of Linear Regression and XGBRegressor, achieving **80%** accuracy

Walmart Global Tech Apr 2021 - Jun 2021
Data Science (NLP) Intern | Intl Global Sourcing team | Awarded a ***Return Offer*** *Bengaluru, India*

- Pre-processed text data and performed data augmentation to solve for class imbalance using **Snorkel's** transformation functions, leading to **5%** increase in training data and **2%** increase in model performance
- Employed SVM, Multinomial Naive Bayes, **Voting classifier**, and **FastText** to perform multi-class text classification. Used **fuzzy match** scores as an evaluation metric & achieved **1%** performance boost over the company's existing model

Projects

MeetMinder App | *Google Vertex AI, Large Language Models, WebVTT* | Jul 2023 - Aug 2023

- Created an application on GCP using **text-bison** LLM to generate meeting's agendas from previous email thread, to live track agendas covered using the zoom transcripts, and to provide minutes of the meet to all meeting attendees

CNN with Layered Boosting & Selective Sampling | *PyTorch, SciPy, CV* | Sep 2021 - Nov 2021

- Implemented the paper Learning to count using CNN boosting by Elad Walach et al. for object counting in images
- Employed Layered Boosting, which adds CNN layers iteratively, and Selective Sampling, which mutes very low or very high error samples, increasing counting accuracy by **20%** and reducing training time by **50%**

Achievements

- **Patent**: **Dynamic Selection of Customized Blocks for Optimized Query Generation*** *Rishabh Thapliyal, Hemant Sharma, Somedip Karmakar et al.* **filed and under legal review with Walmart*
- **Certifications**: GenAI with LLMs, NLP with Tensorflow, Computer Vision, Introduction to Deep Learning, NLP
- **Leadership/Team Work**: IIT Bombay Cricket Team Captain, Department Academic Mentor, IITB Placement Team