KABIR JAISWAL

New York, New York | kabirjaiswal900@gmail.com | 980-446-4489 | github.com/kabir12345

SUMMARY

A computer engineering graduate from New York University with a passion for machine learning and over 4 years of **full-stack experience**. Adept at building robust machine learning solutions in **production environments**, deploying end-to-end ML & **data pipelines**. Actively contributing towards open-source **Generative AI** and **MLOps platforms**. Proficient in **agile** environments and collaborating with cross-functional teams.

SKILLS

- ***ML Tools & Frameworks*:** PyTorch, Hugging Face, TensorFlow, Scikit Learn, SQL & NoSQL, FastAPI
- ***DevOps:*** Terraform, AWS/GCP, Git & GitHub, Docker, Jenkins
- ***MLOps*:** ClearML, MLflow, Apache Airflow, Snowflake
- ***Languages:*** Python, R

EDUCATION

New York University | New York City, NY | August 2022 - May 2024
Master of Science (MS) in Computer Engineering

University of Petroleum and Energy Studies | Uttarakhand, India | July 2018 - May 2022
Bachelor of Technology (B.Tech.) in Computer Science (Concentration: Artificial Intelligence and Machine Learning

WORK EXPERIENCE

Machine Learning – BrandGuard AI | New York, NY — January 2024 - May 2024

- Developed a **custom vector embedding model** for brands with a **500 MFLOPs compute budget** at inference.
- Refined **BrandGPT**, a **RAG** application, with **Cohere's re-ranker**, to improve inference by 3.5-4X over 90th percentile of all queries.
- Spearheaded the **greenfield production** of the **MLOps pipeline**, reducing deployment cycle TAT by 30%.
- Benchmarked LLMs using automation scripts and **Baserun** implementing **OpenAI/Ragas evals** to decrease **cost** and **hallucinations** by 5x.

Machine Learning – Nova AI, Inc. | New York, NY — September 2023 - December 2023

- Launched **synthetic data generation** pipeline for brand ads, using **stable diffusion** models reducing asset creation time by 40%.
- Conducted **throughput analysis** on tools like **MLflow** and **ClearML** using **AWS** to understand QPS limitations (10k reqs/sec)
- Deployed MLflow on **GCP** and **Terraform**, integrating 3 essential components **CI/CD**, **monitoring**, and **versioning.**
- Evaluated **data labeling** solutions (e.g **Human Signal, V7 Labs**) to improve the efficiency of training and testing cycle.

PROJECTS

Spatial Sense | NYU Wireless Labs | Github | January 2024 - May 2024

- Developed a navigation aid for the visually impaired using **edge** devices like the **Qualcomm Gen 2** reducing inference time by 60%.
- Integrating Segment Anything **(SAM)** & Depth Anything **(DAM)** to semantically segment scenes for real-time **obstacle detection**.
- Employing **PEFT quantization** for **on-device** deployment of **LLaVA-1.6** (2-bit quantised) to query segmented object definitions.

Qualitative Analysis of Quantization Techniques for Text Summarization | Github | *December 2023*

- Optimized **Mistral-7b** and **Llama** on CNN **text summarization** dataset boosting BLEU score by 5% (Mistral-7B) and 7% (Llama).
- Leveraged **LORA** and **IA3** to quantize Mistral-7B and Llama models for **on-device** deployment.

GoDesigner | Github | November 2023

- Developed a design recommendation tool with **Flask** and **CLIP embeddings** for precise furniture match based on cosine similarities.
- Utilized **SAM** and **Google Shopping API**, achieving an **F1 score** of 0.92 for accurate product recommendations.

Local Citation Recommendation Systems | Github | *June 2023*

- Optimized embedding models and similarity search techniques to **recommend citations** with a +5% accuracy over SOTA model.
- Utilized **OpenAI**, **Cohere** embeddings with **LangChain's LLMChain Extractor** for citation recommendation improvements.

PUBLICATIONS AND OPEN-SOURCE

Fog Computing Concepts, Frameworks, and Applications-Book | Publication
Co-Author and Academic Research Assistant (University of Petroleum and Energy Studies)

- Chapter 1 - Fog Computing Present & Future
- Chapter 4 - Application of Machine Learning in Fog Computing

Apache Airflow Operator for KDB Integration | Github | PyPi

- Architected and developed an **Apache Airflow** operator for KDB using Python and the Astronomer open-source framework.