Last Updated on July 12th, 2024

Shaik Abdur Rahman Nawaz

github.com/abdurnawaz linkedin.com/in/abdurnawaz abdurnawaz.2011@gmail.com

Experience/Projects

BNY Mellon July 2022 - Present

Software Development Engineer

Eliza (Internal AI Platform)

- Developed an API gateway for LLM access to internal developers, added internal authentication, authorization with AES-GCM encryption to the payloads. At the time of writing, the gateway serves **100,000+** requests daily.
- Built & deployed several LLMs, VLMs, TTS & STT models on-premise for the enterprise using frameworks like vLLM, TensorRT, Triton.
- Created & maintaining a widely used internal Python SDK which allows developers to integrate all APIs from the internal AI platform & some client utilities including authentication/authorization, STOMP & internal cloud storage. The SDK is being used by about **60%** of the teams working on GenAI
- Implemented & contributed to several GenAI features as services/platform capabilities for internal teams to avoid reinventing the wheel. Some of them include RAG(HyDE, BM25, RRF etc.), Code execution engine, Agents with internal data sources & APIs as tools.
- I act as a developer advocate for internal teams working on AI/ML and actively participate in organizing internal AI Hackathons, Expos & Roadshows to help increase AI awareness & adaptability.

RFP Automation

- Contributed to the backend of the first GenAI application in the company that helps the operations team in creating drafts of client RFPs.
- Reduced the loading time of indexes in LlamaIndex from **20mins to 2mins** by adding custom code in the library that skips the need of deserialization of objects after loading indexes.
- Successfully scaled this application to be used by one of the Line of Businesses(LOBs) internally and it is set to expand across the bank's various LOBs.

BNY Mellon May - July 2021

Software Engineering Intern

- Wrote unit tests, improved the code quality and security of various components of an internal ETL application by following the standards of SonarQube.
- Developed a Java application for reconciliation of data which the firm was moving from a third-party vendor to an in-house application.

Personal Projects

Real time Facial Recognition

- Developed a face recognition application that recognises known faces from RTSP camera streams in real time using YOLOv8, ARCFace+ResNet and ChromaDB.
- Designed and implemented a Master/Worker architecture with intelligent load balancing to process multiple video streams utilizing GCP's capabilities.
- Ported the entire codebase from Python to C++ to improve performance which includes building HTTP server, video processing and building the arcface-resnet architecture in CUDA C++.

Education

Indian Institute of Technology Hyderabad July 2018 - May 2022

Bachelor of Tehcnology in Electrical Engineering *GPA: 8.87/10.0*

Skills

Languages	C++, Python, Java, SQL,
Tools	Git, Unix Shell
Frameworks & Other Skills	GCP, Azure, Pytorch, Spring boot