Huzefa Siyamwala

Machine Learning AI Engineer

Interest Areas

Natural Language Processing, Applied Research, Generative AI,
MLOps, LLMOps, Data Quality
A/B Testing, Realtime Inference, Unsupervised Learning

Contact

Phone

408-256-0245

Email

huzefa.siyamwala89@gmail.com

LinkedIn

https://www.linkedin.com/in/huzefasiyamwala

Education

2015

M.S, Computer Engineering
San Jose State University
GPA: 3.9

2011

B.E, Electronics and Communication
Dharmsinh Desai University, India
GPA: 3.7

Frameworks / Tools

- Ray
- MLFlow
- Kubernetes
- Langchain/deepset
- Kafka
- Redis
- Google Cloud
- Tensorflow serve

Languages

- Python
- Nodejs

Datastores

- ElasticSearch
- Weaviate
- Cassandra
- Mysql

Libraries

- Ludwig
- Hugging Face
- Rasa NLU
- Spacy
- Duckling

Experience

2022 - Present

[24]7.ai

Staff Software Engineer, Applied NLP

- Lead Developer on Implementing Declarative ML Platform with pluggable modules for data ingestion. data annotation and data quality.
- Implemented Model LifeCycle Management: Implemented end-to-end framework / workflow for model tuning, model training, model serving and model observability.
- Second pass classifier: Able to consult LLM for intent prediction / slot filling by providing context in dynamic prompts and in context learning with few shots examples
- Scalable and flexible serving architecture reducing research to production time < 1 day
- Currently focused on Self hosted LLM within [24]7.ai infrastructure
- Primary Focus
  - Intent Prediction engine using SOTA LLM (OpenAI, Llama3, Mistral, MPNet)
  - Multi Lingual Support

2015-2022

[24]7.ai

Senior Software Engineer

- Implemented continuous pipeline for human data annotations and feedback pipeline for model training at scale
- Developed Modeling Workbench that enables internal teams to seamlessly build, deploy, and train machine learning models with auto fine tuning support & drift detection
- Worked on content rendering for Apple Business chat and Google Business Messaging

2014

Samsung Semiconductor Inc.

Performance Architect Intern

- Workload characterization: Running TPC-C (OLTP) and TPC-E (Trading) / TPC-H (BI) Benchmarks on Samsung SSDs.
- Prepared Machine Learning model for predicting system performance based on Linux Kernel Configs/MySQL parameters.

2013-2014

San Jose State Research Foundation

Research Assistant

- Used ARM Deassembler to perform statistical analysis of ARM instructions for detecting user to kernel mode transition in resource constrained systems

Publications

- Performance analysis of NVMe SSDs and their implication on real world databases
  - DOI: 10.1145/2757667.2757684
- Identifying Malicious Metering Data in Advanced Metering Infrastructure
  - DOI: 10.1109/SOSE.2014.75