Tony Zhao

tony.ytzhao@gmail.com 🚀 tonyzhao.org 🚀 github.com/cooliotonyio 🚀 linkedin.com/in/tony-zhao

Experience

Software Engineer, ML Infrastructure @ TruEra *Jan 2022 – Present*

- Engineer API/SDK to automate instrumentation, ingestion, and observability of LLM applications
- Lead model ingestion, containerization, and scalable compute strategy on Kubernetes w/ BentoML and Ray
- Orchestrate streaming and batch data ingestion flows w/ Temporal, S3, Flink, Kafka on cloud infrastructure
- Build public performant API communication service in Go for serving both gRPC and REST consumers

Software Engineer, Data Infrastructure @ Ejenta *Feb 2020 – Jan 2022*

- Worked in Scala, Python to create data pipelines as well as interfaces to visualize intelligent systems
- Operated full-stack to build RESTful microservices/interfaces in Java, Scala, Node, and React
- Analyzed health data to inform and design systems to automate customized care plans for patients

Machine Learning Engineer Intern @ Spotify *May 2021 – Aug 2021*

- Implemented machine learning recommendation systems to surface tracks to new users using Python
- Leveraged existing vector embeddings to build models purposed for no-data/low-data contexts
- Redesigned model pipeline to improve latency, simplify dependencies, and ensure user coverage

Software Development Engineer Intern @ Amazon *May 2020 – Aug 2020*

- Created ETL data pipelining in Python with AWS Lambda, State Machines, and Glue
- Built non-blocking validation checks for machine learning model recommendation
- Integrated asynchronous modular functionalities into existing machine learning systems

Education

- **Masters of Science,** *Electrical Engineering and Computer Science*
  UC Berkeley, College of Engineering
- **Bachelors of Science,** *Electrical Engineering and Computer Science*
  UC Berkeley, College of Engineering

Expertise

Programming Languages

Python, Java, Go, Scala, JavaScript/TypeScript, SQL, HTML/CSS

Cloud Infrastructure

Kubernetes, AWS, Helm, GCP, Docker, gRPC, OpenTelemetry, Temporal, RESTful design

Data Infrastructure

Spark, Ray, Kafka, Hive, Flink, MongoDB, Temporal, Trino, S3, Protobuf, Parquet

Machine Learning

BentoML, TensorFlow, PyTorch, HuggingFace, Scikit-learn, SageMaker, Anaconda, LangChain, LlamaIndex, MLflow

Areas of Experties

ML infrastructure, ML Ops, Data infrastructure, Data & Model ingestion/storage/retrieval, ML monitoring, Multi-modal ML, Embedding spaces, Recommendation systems, API design, LLM instrumentation/observability