Tim Xia

Full-stack / Distributed Sys / ML Infra

https://github.com/ttback https://ti3x.github.io/ https://www.linkedin.com/in/timxia

Experienced engineering leader with hands-on GenAI experience and 10+ years of launching and managing complex projects with cross-functional teams across geographic locations

Education

Georgia Institute of Techonlogy
Bachelor of Computer Science
Graduated with honor(2011)

Skills

Backend: Python, Golang, Java, Node.js, Rust, PostgreSQL, SQS/SNS, Redis, Kafka, Mongodb, Django, ElasticSearch, Langchain, FastAPI, Qdrant, Faiss, Lancedb

Frontend: React/Redux/React-Native, Typescript, jQuery, Bootstrap/MaterialUI, Web Components, Leaftlet, D3, Graphql, PhoneGap

Cloud & Ops: AWS, GCP, Docker, Kubernetes, Git, JenkinsCI, Datadog

ML/Research: Web-AR, Object Recognition, NLP, OCR, webassembly, MLOps, Tensorflow/Pytorch, TensorRT-LLM

Certifications and Courses

Neural Networks and Deep Learning
(https://coursera.org/share/b2f9f45d6a989506fbe1936c710bfcd1)

Notable Projects

LLM Agent with Medical Ontology
Reduce LLM hallucinations with both model fine-tuning and ontology knowledge graph

Relevant Prior Search
Created semantic search for medical reports based on multi-modal vector data using clip and vector dbs such as lancedb, qdrant, and faiss

Patent Pending

US Patent Application 63598037 “AI-Assisted Medical Image Report Generation Display and Workflow,” November 10, 2023

References

Available upon request

Experience

Sirona Medical | AI team **June 2022 - Present**
Tech Lead | Senior Software Engineer

- Supported CTO’s product demos that 2x company ARR via collaborative development
- Designed and developed a proprietary MLOps and LLM Ops platform that includes model serving, feedback gathering, LLM Evals and prompt engineering to support product prototyping and ideation for radiology workflows powered by generative AI
- Successfully shipped low-latency and high quality(90% met user preference) LLM-powered features to Sirona Workspace and added multi-modal integrations with LVM
- Developed multiple GenAI workflow features with Langchain, RAG/Vector db, model fine-tuning, and agents that increased AI adoption and increased case volume to millions
- Created and fine-tuned speech-to-text models using synthetic data for speech assistant
- Developed voice assistant prototype with RASA in combination with LLM and vector db
- Led MLOps integration with OpenAI, AWS Sagemaker, GCP Vertex, Anthropic and fine-tuned llama models, along with HuggingFace, TensorRT-LLM, Langchain and Ray
- Experience building ML Infra with Kubeflow, MLFlow, Triton and ML model compilers

NYTimes | Search and Content-Tagging team **Nov 2018 - June 2022**
Senior Software Engineer

- Led Search team of 5 on the easy-to-use and user-centric interface development of internal Timespast Research portal redesign and new NYT Search features with React, Redux, Graphql, Web Components, and Typescript
- Applied ML by creating a streaming image ingestion pipeline for internal NYT image semantic search with Tensorflow, CNN and transformers and multiple vector dbs
- Worked on Elasticsearch optimizations for relevance and improved deployment and monitoring capabilities with Kubernetes, Terraform, and Clojure
- Built ML pipelines and React UIs for collaborative content tagging, NLP topic research features, and precision/recall analysis in collaboration with taxonomists, newsroom, and 3rd-party vendors on the Semantic API platform for content curation and classification
- Partnered with designer, newsroom, analysts, and QA to improve user engagement and experience and increased search result CTR to 32%-48%

NYTimes | Knowledge Management team **Nov 2016 - Nov 2018**
Senior Software Engineer

- Served as technical gatekeeper performing technical reviews for NYT's archive schema standardization and architecting large-scale data pipelines and CI/CD practices
- Completed massive data ETL pipelines in Python and Golang for Archive migration and modernization and various tooling for Protobuf-based data schema for NYTimes' new streaming publishing pipeline built on top of Kafka

NYTimes | R&D and Discovery team **Aug 2013 - Nov 2016**
Software Engineer

- Created Timesmachine, a progressive and interactive web archive microfilm reader that provides access to 171 years of archive and added 10k+ new subscriptions each year
- Collaborated with UX designers to build a collective product vision for user engagement
- Built TimesMachine Publisher, an advanced Python newspaper map tiler application customized for robust parallel execution at scale on Amazon MapReduce clusters. The latest Timesmachine Publisher can tile 100 tera-pixel newspaper scans/4 minutes on 460 8-core clusters for 3.6+ million scans
- Mentored and managed 3 interns to ship production and research features

Fidelity Investments | Personalized Portfolio Team **July, 2011-Aug, 2013**
Associate Software Engineer

- Gathered requirements from portfolio managers and analysts to create asset allocation tools with large-scale financial data for Fidelity Personalized Portfolio VIP customers
- Built a real-time collaboration dashboard with Node.js, Socket.io, Bootstrap and Ruby