Yike(Eric) Tan

http://likegiver.github.io

Email : yiketn@gmail.com
Mobile : +1-412-583-0350

Education

Carnegie Mellon University Pittsburgh, PA

Master of Science in Artificial Intelligence Engineering - Information Security; GPA: 3.78/4.00 *Dec. 2025 (Expected)*

- **Coursework**: LLM Systems, LLM Methods and Applications, AI Systems and Tool Chains, Computer Systems

University of International Business and Economics Beijing, China

Bachelor of Engineering in Data Science and Big Data Technology; GPA: 3.70/4.00 *Aug. 2020 – Jul. 2024*

- **Coursework**: Operation Systems, Data Structure and Algorithm, Natural Language Processing, Big Data Analysis

Experience

Instance Creator LLC Sunnyvale, CA (Remote)

Software Engineer Intern *May 2025 - July 2025*

- Spearheaded a vector-based semantic matching system using **Qdrant**, boosting matching accuracy to over 80%
- Developed a multi-agent automation platform with **AutoGen** and gpt-4o to orchestrate complex application workflows, reducing manual effort by 85%; automated resume customization and submission tasks via **Playwright**.
- Designed and implemented a scalable microservices backend using async **FastAPI** to support over 1,000 concurrent users; optimized performance by introducing a **Redis** caching layer, reducing external API latency and call frequency by 80%.

Epoching AI Beijing, China

Software Engineer Intern *Jul. 2024 - Aug. 2024*

- Achieved a 10x inference speedup and 30% accuracy boost for a watermark removal service using **NVIDIA TensorRT**.
- Developed a data generation and augmentation pipeline, improving model robustness against semi-transparent watermarks.
- Containerized the service using **Docker** and deployed it on **Cloud ECS**, implementing **Prometheus** for real-time monitoring and ensuring high availability to handle 500+ concurrent requests.

Ytell Network Technology Co., Ltd. Beijing, China

Software Engineer Intern *Sep. 2023 - Mar. 2024*

- Developed a high-performance multimodal RAG chatbot with **FastAPI**, increasing inference throughput by 5x using **Flash Attention 2** and cutting retrieval errors by 40% through a hybrid search strategy.
- Automated the processing of 1,000+ daily orders with a data pipeline using **Apache Airflow** and **MongoDB**, eliminating over 5 hours of weekly manual work for each store manager.

Projects

ImaginAItion: AI Literacy Game - CMU HCI *July 2025 - Present*

- Engineered a real-time multiplayer AI literacy game with a **React (TypeScript)** frontend and **FastAPI** backend; leveraged **Socket.IO** over WebSockets to ensure low-latency (<150ms) state synchronization for 50+ concurrent players.
- Implemented a stateful backend game loop to manage turn-based player progression, integrating the gpt-image-1 API for real-time, dynamic image generation.
- Orchestrated the full-stack application with **Docker Compose** for reproducible local development and deployed the containerized services to **Amazon ECS** for a scalable production environment.

Silent Supporter: Scalable AI Multimodal Therapy Platform - Tsinghua AIR *June 2025 - Aug 2025*

- Built a scalable therapy platform on a **Node.js** backend, reducing real-time generation latency by 80% via **WebSockets**; utilized **PostgreSQL** for primary data storage and **Redis** for session caching to improve performance.
- Achieved targeted music style transfer by fine-tuning the InspireMusic 1.5B model on a custom-built dataset, and developed a **WebGL** engine to dynamically visualize conversational emotion in real-time.
- Designed a low-latency, event-driven architecture using **RabbitMQ** for asynchronous, non-blocking multimodal generation; separately optimized the core AI model's inference speed by 3x through **ONNX Runtime** graph optimization

MovieRec – End-to-End Recommendation System *Feb 2025 - Mar 2025*

- Architected a full-stack recommendation platform, integrating a **Vue.js** frontend with a backend based on **Flask API** and **SpringBoot**; Leveraged **MySQL** and **Redis** to handle the data persistence layer.
- Trained Random Forest model using **Scikit-learn**, conducted offline evaluation with train-validation splits and online evaluation with telemetry data to assess model performance; managed A/B testing experiments with **MLflow**.
- Established a containerized MLOps workflow, simulating a high-availability architecture with **Minikube** and implementing a **Jenkins** CI/CD pipeline with **Prometheus/Grafana** for automated deployment and real-time monitoring.

Self-LLM: Open-Source LLM Deployment Guide (23.2k Stars) *May 2024 - Jun 2024*

- Co-authored and maintained a leading open-source guide simplifying LLM deployment, which led to its feature in the keynote presentation at **Google I/O Connect China 2024.**
- As a core contributor, I authored key tutorials on **GLM4** deployment with **vLLM**, **LangChain** integration, and **LoRA** fine-tuning to simplify the learning curve for developers.

Programming Skills

Programming Languages: Python, C, CUDA, C++, Java, JavaScript, SQL, R
Packages: PyTorch, TensorFlow, Triton, Jax, SparkML, NumPy, Pandas, React, Node.js, Django, Flask
Tools: Git, Docker, Kubernetes, Kafka, Neo4j, Amazon Web Service, Google Cloud Platform, Unix, LaTeX