



RUILIN JIN

Mobile/WeChat: (+86)13552110415 | sam.ruilin@hotmail.com |  | 

Education

Case Western Reserve University

Master of Science in Computer Science | Artificial Intelligence Track | GPA 3.8/4.0

Sep. 2023 - May 2025

Cleveland, OH

Rensselaer Polytechnic Institute

Bachelor of Science in Information Technology & Web Science | Minor in Philosophy

Sep. 2016 - May 2021

Troy, NY

Relevant Coursework: Designing High Performance Systems for AI, Machine Learning and Causal Inference, Deep Generative Models, Natural Language Processing, Computer Vision, Large Language Models, etc

Professional Experience

China Railway Cloud Information Technology Co., Ltd.

Software Engineer | Employee of the Year Award in 2022

Jun. 2021 - Aug. 2023

Beijing, China

- **AI Middle Platform**
 - * Developed a text similarity comparison feature using the BERT model, which achieved a 95% accuracy rate, reducing processing time from 3 hours to 2 minutes.
 - * Applied LoRA and P-Tuning techniques in fine-tuning the ChatGLM-6B model on a private dataset, increasing QA accuracy by 15% and user satisfaction to 84.8%.
 - * Constructed a Kafka real-time data streaming system, optimized the cost of data processing and storage by 60%; the system processes over 20 GB of data daily.
 - * Developed a facial recognition system for construction sites, achieving a 98% recognition accuracy, saving 21,000 yuan in third-party API fees per month.
- **R&D Platform**
 - * Participated in the design of the micro-frontend architecture; built a component library with Vue.js, optimized the implementation with Nginx, Docker, and Jenkins, established coding standards, accelerated development cycle by 40%, reduced implementation time by 50%, served 500,000 users, and completed and delivered 43 projects annually.
- **Marketing Management System**
 - * - Served as the front-end team leader, led a team of 11 people to complete the development of over 100 pages and 8 major modules within three months; applied virtual scrolling and lazyload techniques, which reduced page loading time by 90%; improved product quality and user experience in a data-driven manner.

IBM

Software Engineer Intern

Sep. 2019 - Dec. 2019

Troy, NY

- Participated in the Watson Digital Twin Technology project; designed use case analysis and developed prototypes; implemented RESTful API communication with React.js and GoLang (Gin framework); applied Pandas and Tableau for data analysis to reduce costs and improve ROI.

Research and Selected Project

KV Cache Optimization in Large Language Model Inference

Graduate Research Assistant, CWRU

Jun. 2024 - Present

Cleveland, Ohio

- Designed and implemented KV Cache optimization strategies; used pooling techniques and score-based attention selection mechanisms to effectively reduce cache occupancy.
- Managed the Token generation process, optimized attention distribution, improved the model's inference accuracy and responsiveness.
- Conducted in-depth analysis of the Draft Model; identified and implemented optimization and deletion strategies for KV Cache to enhance model performance.

Generative AI in Education, RAG Agent

Graduate Research Assistant, Weatherhead School of Management, CWRU

Mar. 2024 - Present

Cleveland, OH

- Integrated multiple cutting-edge technology stacks; automated project deployment with Vercel and GitHub Actions, which saved 32 working hours daily and effectively processed over 8,000 requests, significantly improving development efficiency and system stability.
- Developed identity authentication middleware, integrated it with the school's SSO system, and used JWT for identity verification; built an RBAC-based access control system that comprehensively managed interface calls, user permissions, and access control, successfully preventing over 300 unauthorized access attempts.
- Embedded data into Pinecone namespaces, implemented a multi-tenant system, improved data retrieval efficiency, and reduced retrieval time from 45 seconds to 10 seconds.
- Developed the Retrieve Merger feature and optimized the effectiveness of Retriever in LangChain; achieved multi-source queries and provided accurate answers, improved the proxy response accuracy rate for complex multi-source queries by 25%, hence increasing user satisfaction by 20%.

Hello Algo Open Source Project | | 94k Star

Contributer

Dec. 2023 - Present

- Responsible for writing and translating tutorial content, ensuring the materials' accuracy and readability to match learners with different backgrounds.

Technical Skills

Languages: Python, JavaScript, TypeScript, GO, SQL, Java, C++

Database: Pinecone, MySQL, PostgreSQL, MongoDB, DynamoDB

Technologies/Frameworks: PyTorch, MLFlow, TensorFlow, Vue.js, React.js, Next.js, Redux, Flask, FastAPI, Docker, Kafka, AWS S3, Git