Chongwen (Antony) Zhao

MEng in Electrical & Computer Engineering @ Duke
Software Development Engineer, 2026 NG (Starting June 2026)

cz207@duke.edu
(919) 236-9216
linkedin.com/in/antony957
github.com/antony957

Education

Duke University, Pratt School of Engineering Aug. 2023 - May. 2026
Master of Engineering in Electrical and Computer Engineering GPA: 4.0/4.0
Core Courses: Machine Learning, Deep Learning, Compiler Construction

Hohai University, School of Computer and Information Aug. 2018 - May. 2022
Bachlor of Engineering in Computer Science GPA: 4.42/5.0 (87.44/100)
Core Courses: Operation System; Computer Network; Software Engineering; Data Structure

Experience

Scalable Platforms & Distributed Systems

Dispatching and Playback System @ Meituan (Full time) Mar 2022 - Nov 2022

- Led and delivered a large-scale playback system to process and visualize billions of mobility data points, reconstructing dispatch flows and bike trajectories, reducing annual manual workload by ~5,000 hours.
- Built dispatching and scheduling platforms supporting city-level customization and lifecycle management of rules, enabling cross-city scheduling analysis for optimization, significantly reducing manual effort.

Model Experiment Platform @ Meituan (Full time) Jun 2022 - Jun 2023

- Built a scalable simulation platform to evaluate dispatching models under realistic mobility scenarios, enabling the team to simulate and assess 100+ strategies efficiently.
- Developed online A/B testing frameworks to rigorously measure model effectiveness and ensure stable production rollouts of newly trained models.

Data Pipeline for Model Training @ Meituan (Full time) Sep, 2022 - Jun 2023

- Implemented a high-throughput data pipeline to process 10M+ logs per day using Flink and HBase, ensuring scalable storage and efficient retrieval for training workflows.
- Delivered comprehensive training and evaluation datasets to algorithm teams, shortening model tuning cycles by 16% and boosting group ROI by 25%.

Applied AI Engineering

Talking Character App (LLM-powered) @ Simpleway.AI (Part-time, Remote) May 2025 - Aug 2025

- Designed and implemented a scalable real-time voice chat platform (React/Next.js + Flask + PyTorch backend) with WebSocket-based streaming, supporting low-latency interactive sessions.
- Built modular service pipelines with hot-reload model integration, enabling dynamic updates without downtime and improving system reliability in production.
- Collaborated with product and animation teams to integrate rendering modules, delivering a seamless end-to-end user experience across thousands of daily active sessions.

AI Resume Matching System @ Simpleway.AI (Gap Year, Full Time) Jan 2025 - May 2025

- Designed and implemented an end-to-end resume processing service with OCR and embedding-based retrieval, building a scalable backend pipeline capable of processing 50K+ resumes daily with average response latency under 200ms.
- Developed a semantic matching engine to align resumes with job descriptions, and integrated the service into the clients recruitment workflow, reducing manual screening workload by 80% and improving system throughput by 35%.

LLM Safety Engineering @ Premilab (Gap Year, Full time) Jun 2024 - Jan 2025

- Designed and implemented security control modules for large language model inference services, strengthening protection against adversarial inputs and jailbreak attempts.
- Evaluated robustness and performance across multiple distributed deployments, reducing jailbreak attack success rate from 70% to 5% while maintaining stable throughput.

Skills

Backend & Data: Python, Java, C++; Web frameworks (FastAPI, Spring Boot); Data stream processing (Flink); SQL/NoSQL databases (HBase, Hive, MySQL); distributed systems (zk, Nginx)
Infrastructure: Docker, CI/CD, Redis, Message Queues (Kafka/RabbitMQ), RPC (Thrift), Cloud deployment
Frontend: React, Tailwind CSS, Next.js; Real-time communication (WebSocket)
AI/ML: Transformers, LLM finetuning (LoRA, PEFT), RAG