Xin Cui

• chloexincui@gmail.com • arriettyc.github.io

Work Experience

Aug 2025 – Current	**Founding Engineer (Stealth) / Research Lead (Tzafon AI),** `San Francisco, CA US` Conducting applied research on agentic coding workflows. Previously led post-training strategy for computer-use agents at Tzafon AI.
Jul 2024 – Jun 2025	**Research Engineer,** `Character AI, Menlo Park, CA US` LLM Post-training team
May 2018 – Mar 2024	**Senior Software Engineer,** `LinkedIn, Mountain View, CA US` Retention AI Platform, Job Search and Recommendation team
Jun 2015 – Dec 2016	**Full-Stack Software Engineer,** `Indeed, Tokyo Japan` SEO, Page Ranking team
July 2014 – Sep 2014	**Software Engineer Intern,** `CITIC Securities, Shanghai China` Quantitative Trading Service team

Projects Experience

Current / Tzafon

Agent-Driven Development Research (Stealth)

• Evaluating the limits of SOTA coding agents (e.g., Claude, Gemini) in building full-stack applications, focusing on failure modes in long-horizon reasoning and self-correction loops.

Post Training Lead: Computer Use Agent (Tzafon)

• Unified SFT Architecture: Redesigned the fragmented training codebase into a unified, modular SFT pipeline for Qwen VLMs, eliminating redundancy and reducing technical debt to enable rapid iteration.
• Infrastructure Scaling (100x): Engineered the data ingestion layer to scale from single-gigabyte datasets to ∼100GB of multimodal trajectories, unlocking high-fidelity fine-tuning capabilities.
• Multi-Modal Evaluation Suite: Built a versatile evaluation engine supporting OSWorld and custom GUI screenshot benchmarks, establishing a standardized protocol for measuring visual grounding and action execution.

Character AI

OSS Fine Tuning

• Large scale distributed SFT training with ray and deepspeed on the Qwen3 dense text model series

Data Curation in Preference Alignment

• Curated training data for Direct Preference Optimization (DPO), including privacy-preserving chat clustering to support pluralistic alignment. Designed scoring and filtering strategies to select safer and more balanced preference pairs. Fine-tuned language models using iterative DPO and its variants, monitored training via LM loss, KL divergence, and preference margin metrics, and evaluated model performance through online A/B experiments.

Safety Modeling

• Engineered an end-to-end transformer-based safety classifier, partnering cross-functionally to translate safety policies into technical ontologies and red-teaming protocols. Automated the data processing pipeline to scale ingestion from megabytes to gigabytes, utilizing LLM distillation, reject sampling, and human-in-the-loop majority voting. This pipeline yielded a 10x increase in balanced synthetic training samples, directly driving an approx. 8% improvement in precision with sustained recall, validated through rigorous offline evaluation and online A/B testing.

Tech-stack – python, spark/k8s, apache beam/Google dataflow, SFT/DPO/pytorch, MySQL/Datalake/BigQuery, WandDB/Hex, Training(Ray/Hydra/Deepseed), Inference(vllm/roller), vscode/git/hex, dbt / apache airflow

LinkedIn

Contextual Next-Best-Action Recommendation Platform and Product

• Built contextually triggered cross-pillar (feed, jobs, edge building etc) action recommendations push notification for infrequent LinkedIn members, redesigned Next-Best-Action comms served as a platform and allows partner team to onboard, collaborate with Retention AI team. Impact site-wise +0.1% WAU.

Near realtime audience targeting

• Improved A/B experiment member targeting efficient from offline manner to near realtime, improved data freshness from one day delay to P99 5-10 seconds. Since the A/B experiment platform is serving as one of the largest QPS in LinkedIn, the new pipeline is able to serve high data freshness with low latency.

New search index for job seeker and recruiter

• Led and implemented fair chance employer filter and had multiple experiences on backend index building across LinkedIn's major products – Jobs and Recruiter. Brought Product team's novel search experience to reality.

Job's Data Streaming Platform

• Built data streaming pipeline with Apache Samza to standardize implicit job data/features. With combination of rule base and fastText model, able to achieve the data extraction with precision to 97% across all targeted benefits with avg recall 67%

Tech-stack Java, Scala, Python, Spring, Hadoop, Key-Document DB(Espresso), Kafka, Restful APIs, Apache Samza, Apache Lucene

Indeed Tokyo

Enrich Companies' Pages with Their Instagram Photos

• Link companies' official Instagram accounts to their Indeed accounts and show the latest Instagram photos on their Indeed company pages.

• Brought Hackathon project to Product. Two engineers and one PM hackathon project that aiming to solve less enriched company showcase page by integrating with Instagram account and display most recent instagram photo posts.

Improve Searching Engine Optimization(SEO) Strategy

• Collaborated with XFN (Marking Strategy, PM) teams to implement and experiment various SEO features to drive organic traffic to the site. Worked as a full-stack software engineer.

Tech-stack Java, ReactJS, Spring, MySQL, Google Closure Library, CI/CD

CITIC Securities

Quantitative Trading Data Serialization Support

• Implemented a dynamic serialization declaration mechanism and added support for WCF-standard DataContract and DataMember attributes, enhancing compatibility within the existing data exchange framework.

Tech-stack C#, Protocol buffers.

Contest

2010 **Bronze Medal** `35th ACM-ICPC Asia Regional Programming Contest`

Education

2012 – 2015 **Fudan University**, Computer Science, Master Degree
Shanghai Key Laboratory of Intelligent Information Processing | Advisor: Prof. Shuigeng Zhou

Focused on applying Latent Dirichlet Allocation (LDA) for single-cell RNA-seq (scRNA-seq) analysis.

2008 – 2012 **Northeast Normal University**, Computer Science, Bachelor Degree
National Scholarship, Ministry of Education of China