Ayan Mao

Palo Alto, CA
206-673-6739 mayforsde@gmail.com linkedin.com/in/ayan-mao 98may.github.io

About Me

Software Engineer with 2.5 years of experience across X/xAI, Google, and Meta, building large-scale **distributed systems** and **applied AI** infrastructure for real-time ads.

Proven ability in high-QPS production ownership, latency reduction, and LLM-powered retrieval and tooling that improves matching quality and system interpretability.

Experience

xAI **March 2025 – Present**
Member of Technical Staff @ Applied AI / Ads Team — Java, Scala, Python *Palo Alto, CA*

X **Feb. 2024 – March 2025**
Software Engineer I/II @ Ads Team — Java, Scala, Python *Palo Alto, CA*

- Maintained and optimized high-throughput distributed ads systems serving **hundreds of thousands QPS** (tens of billions of requests per day), achieving **99.95% availability** as **one of three primary on-call engineers**
- Led migration from legacy service architecture to next-generation sequence ranking model with two junior engineers to eliminate technical debt and achieve **80ms+ (p95) latency improvement** and website conversion gains
- Built Grok-based candidate retrieval system by integrating **Grok LLM for semantic user and content understanding**, developing real-time inference pipeline with efficient caching that became **one of the highest-performing retrieval sources**
- Created internal **web tool for prompt tuning and result visualization** for Grok-based ads and dynamic product ads, significantly improving matching quality and system interpretability
- Integrated Google AdMob demand into X's ad serving stack working with iOS and Android teams, building scalable backend infrastructure that grew annual revenue from **xM to x00M** within 5 months
- Optimized ad fatigue quality by funnel analysis and hybrid impression checking that increased **user engagement by 1.3%** and **user active seconds by 0.12%**

Internships

Google **Aug. 2022 – Nov. 2022**
Software Engineering Intern @ Mobile Platforms — Java, Lua, GCP *Mountain View, CA*

- Built full-stack prototype for 3D virtual environment integration, developing a Lua SDK for serving dynamic content in immersive platforms (e.g., Roblox), which scaled to become a production project with multiple full-time engineers
- Optimized mobile ads' close button and released it for billions of Android users using Android Studio

Meta **May 2022 – July 2022**
Software Engineering Intern @ Infrastructure — C++, Python *Menlo Park, CA*

- Enhanced testing infrastructure by automating A/B test orchestration across development and production environments, building commit-level performance diagnostics that improved debugging efficiency for backend services

Microsoft Research Asia **Oct. 2020 – June 2021**
Research Intern @ Systems Research Group — Python, TypeScript *Beijing, China*

- Contributed hyperparameter optimization algorithms to NNI (Neural Network Intelligence, 13.2k stars AutoML framework), implementing auto-tuning strategies that increased downstream training effectiveness

Technical Skills

Languages: Java, Python, Scala, C++, Rust, Lua, TypeScript
Backend & Tools: Distributed Systems, Microservices, Real-time Ads Serving, High-throughput APIs, MySQL, Hadoop
Applied AI & Platform: LLM Integration, Model Serving, AWS/GCP, Docker/Kubernetes

Education

Northeastern University **2021 – 2023**
Master of Science in Computer Science *Seattle, WA*

Zhejiang University **2017 – 2021**
Bachelor of Engineering in Computer Science *Hangzhou, China*