Shanli Xing

@ shanlx@uw.edu | GitHub | Website

Education

University of Washington — Seattle, WA
Bachelor of Science & Honors Program; Computer Science; GPA: 3.73/4.00 — *Sep 2022 – Jun 2026 (Expected)*

Experience

Paul G. Allen School of Computer Science & Engineering — Seattle, WA
Research Assistant — *Jan 2024 – Current*

- Advised by *Prof. Luis Ceze* and collaborating with Ph.D. student *Zihao Ye* on building efficient machine learning systems.
- Experience in GPU programming (Cuda, Triton, Cutlass), Pytorch, and LLM inference optimizations.

Projects

FlashInfer | *Github*

- A library and kernel generator for Large Language Models that provides high-performance implementation of LLM GPU kernels.
- Co-designed the dual-pivot sampling/renormalization algorithm and implemented the kernels. Built the SM90 group GEMM kernels and Python interface. Involved in other library maintenance like addressing community requests.
- Authored the technical blog post *Sorting-Free GPU Kernels for LLM Sampling*, introducing the motivation, design iterations, and implementation of our sampling algorithms; received significant community traction, including over 30k views on X.

Radicle | *Github* | Course Project of *Systems for ML*

- An inference framework implementing Prefill/Decode disaggregation on distributed clusters. Featureing runtime elasticity-expanding and shrinking the cluster on the fly, node multiplexing through dynamic role reassignment, and fault tolerance through per-step timeout monitoring.
- Proposed a controller-actor architecture, implemented with Ray. The central Controller actor orchestrates Prefill and Decode actors to manage asynchronous prefill batching, KV-cache handoff, and continuous token streaming back to clients.

ChatFeedback | *Github* | Course Project of *Data-centric ML*

- Inspired by *Chatbot Arena* and *OpenAssistant*, aimed to sustainably crowdsource high-quality, scalable, and diverse human preference data for LLM alignment.
- Designed a user incentive system incorporating compute points and gamification to motivate contribution, and a confidence score mechanism to ensure data quality.
- Implemented a full-stack web application with a responsive and attractive UI using React.js, Next.js, and Tailwind CSS. Utilized Vercel for serverless CI/CD, with cloud PostgreSQL and Redis for persistent storage and caching.

ReedIsland | *GitHub*

- An Android forum client, structured following MVVM architectural pattern.
- Developed based on an open-source application and significantly modified to incorporate the different features and APIs of the new forum, including post retrieval, display, creation, and interaction; user management, bookmarking, browsing history, voting, and custom scripts.
- Updated and refined the UI to support new features. Performed extensive upgrades and refactoring of project dependencies. Contributed approximately 10k lines of meaningful code changes.
- Reached 1000+ active users according to Visual Studio App Center.

Relevant Coursework

Graduate: Deep Learning, Natural Language Processing, Systems for ML, Distributed Systems, Data-centric Machine Learning, Ethics in AI, (Current) Advanced Machine Learning

Undergraduate: Computer Vision, Systems Programming, The Hardware/Software Interface, Programming Languages, Software Design & Implementation, Foundations of Computing (Discrete Math & Probability), Linear Algebra, Differential Equations, (Current) Computer Graphics