# Shao Wang

shaowang2002@gmail.com | github.com/Electronic-Waste

## Research Interests

My research interests lie at the intersection of machine learning and computer systems. I'm currently focusing on building efficient and large-scale systems for emerging ML workloads like LLMs and AI agents.

## Education

**Shanghai Jiao Tong University** | *Computer Science M.S.* 2024.09 - Now

- Cumulative GPA: 3.81/4.00
- Award: Optiver Outstanding Scholarship, First-Class Academic Scholarship

**Shanghai Jiao Tong University** | *Software Engineering B.Eng.* 2020.09 - 2024.06

- Cumulative GPA: 3.76/4.30
- Award: Huawei Scholarship, Wenyuan Pan Scholarship

## Research Experiences

**Catalyst Lab, Carnegie Mellon University** | *Research Assistant* 2026.04 - Now

- Advisor: Prof. Zhihao Jia.
- Currently working on an agent serving project.

**Data Service Lab, Shanghai Jiao Tong University** 2025.07 - 2026.04

- Advisors: Prof. Lin Gui and Rui Ren (Telecommunications).
- Initiated ForkKV (in submission) project, which alleviates the memory footprint bottleneck in multi-LoRA agent serving by enabling efficient KV cache sharing across agents. Achieves up to 3× the throughput of SOTA baselines.
- **Independently led the project** from conception to implementation and manuscript composition.

## Publications & Talks

**ForkKV: Scaling Multi-LoRA Agent Serving with Copy-on-Write Disaggregated KV Cache** 2026.04

- ***Shao Wang**, Rui Ren, Lin Gui*
- *In submission, arxiv.org/abs/2604.06370*

**RAG and Fine Tuning with Kubeflow** 2025.11

- *Francisco Javier Arceo, **Shao Wang***
- *Talk at KubeCon + CloudNativeCon NA 2025*

**Streamline LLM Fine-tuning on Kubernetes with Kubeflow LLM Trainer** 2025.11

- ***Shao Wang**, Andrey Velichkevich*
- *Talk at Kubeflow Summit NA 2025*

## Open Source Projects

**Kubeflow** | *WG AutoML/Training Maintainer* 2024.10 - Now

- (Kubeflow has earned **33.2k+** stars on Github and **268M+** downloads on PyPI over the years)
- **LLMOps:** Led the design and implementation of Kubeflow LLM Trainer V2, which streamlines LLM fine-tuning on Kubernetes by: 1) Hiding complex Kubernetes configurations from data scientists via designing CRDs for different personas 2) Unleash data scientists from the burden of manually setting up environment by providing editable workflow templates and a simple yet efficient Python API
- **Mentorship:** Acted as a Mentor for several Kubeflow GSoC'25 projects, leading a three-member team and coordinating cross-time-zone collaboration to support students in accomplishing their project goals.
- **Maintenance:** Participate in the daily maintenance routine for *kubeflow/trainer*, *kubeflow/katib*, and *kubeflow/sdk*

## Industry Experiences

**Databend Labs** | *Cloud Platform Intern* 2024.10 - 2025.04

- **Operator:** Developed an open-source operator for databend on Kubernetes from scratch, providing open-source users with a flexible yet efficient approach to orchestrate databend clusters on Kubernetes.
- **Observability:** Integrated Prometheus for enhanced system visibility on the on-premise cloud platform.

**Goolge Summer of Code** | *Contributor, CNCF - Kubeflow* 2024.05 - 2024.09

- **Push-based Metrics Collection:** Built a push-based training metrics system to replace Katib's sidecar approach, which reduced performance/resource overhead and improved metrics collection latency.
- **SDK:** Developed a simple Python API for metric reporting, minimizing format errors and improving robustness.
- **E2E Testing:** Added e2e tests for hyperparameter tuning in GitHub Actions to enhance CI coverage.

**Bondi Tech Ltd.** | *Infrastrcture Intern* 2024.01 - 2024.04

- **Investigation:** Evaluated time series databases for OLAP analysis of market data and delivered benchmark reports.
- **Quote Storage:** Build a new storage system from scratch on Kubernetes in replace of the current CSV-based storage system, which enhances the performance of data analysis business by 10x times.