Muxin (Murphy) Tian

murphy.tian@mail.utoronto.ca | LinkedIn | GitHub | Google Scholar | Website

EDUCATION

University of Toronto, St. George Campus — Toronto, Canada
Bachelor of Computer Science, Arts & Science Internship Program — Sep 2021 – Jun 2026

SELECTED PUBLICATIONS

* denotes equal contribution

1. **Muxin Tian**, Haoran Ni, Yiyan Zhai, Yangsun Park, Juncheng Yang
   "RouteWise: Latency–Cost Optimization for Multi-Provider LLM Routing". Submitted to Eurosys 2027.
2. **Muxin Tian***, Zhe Wang*, Zhenwei Tang, Blair Yang, Kunlun Zhu, Honghua Dong, Hanchen Li, Guangjing Wang, Jiaxuan You
   "SWE-Bench Mobile: Can Large Language Model Agents Develop Industry-Level Mobile Applications?" **KDD 2026 ADS Track Main Conference**
3. Kunlun Zhu*, **Muxin Tian***, Zijia Liu*, Bingxuan Li*, Yingxuan Yang, Jiaxun Zhang, Pengrui Han, Qipeng Xie, Fuyang Cui, Weijia Zhang, Xiaoteng Ma, Xiaodong Yu, Gowtham Ramesh, Jialian Wu, Zicheng Liu, Pan Lu, James Zou, Jiaxuan You
   "Where LLM Agents Fail and How They Can Learn From Failures".
4. Hannah Liu*, **Muxin Tian***, Iqra Ali, Haonan Gao, Qiaoyiwen Wu, Blair Yang, Uthayasanker Thayasivam, Annie Lee, Pakawat Nakwijit, Surangika Dayani Ranathunga, Ravi Shekhar
   "OasisSimp: An Open-source Asian-English Sentence Simplification Dataset". **LREC 2026 Oral**

RESEARCH EXPERIENCE

Harvard MadSys Lab, Harvard University — Jun 2025 – Present
Research Intern, supervised by Prof. Juncheng Yang

- Led and operated FreeInference end-to-end, owning the architecture, implementation, provider integrations, production operations, and rollout for an open LLM inference platform serving Harvard and MIT campuses.
- Designed and implemented **RouteWise**, a cost-latency router for multi-provider LLM inference that dynamically routes across on-demand API, quota plans, and concurrency subscriptions; reduced API cost by **40%** and P99 latency by **4.5x** in real-provider evaluations. Submitted to EuroSys 2027.
- Proposed **Nimbus**, a hybrid inference system that combines dedicated GPU deployments with elastic serverless APIs to absorb bursty workloads under latency SLOs; reduced bursty TTFT by 11.7x while avoiding serverless-only API costs. Preparing submission to ATC 2026.

ULab, UIUC — Feb 2025 – Feb 2026
Research Intern, supervised by Prof. Jiaxuan You

- Co-led **OpenManus-RL**, an agentic reinforcement learning framework for fine-tuning LLM agents on environments such as ALFWorld, WebShop, and GAIA; integrated benchmarks, unified rollout formats, and migrated training pipelines onto veRL / verl-agent for large-scale experiments.
- Co-developed **AgentDebug**, a framework that analyzes LLM agent trajectories with a modular error taxonomy to localize root-cause failures and provide targeted feedback; contributed trajectory integration, debugging, ablation experiments, and paper writing.
- Initiated and led **SWE-Bench Mobile**, an industry-level benchmark for mobile app development agents, reconstructing real-world iOS/Android workflows from PRDs, Figma designs, and production codebases; led task formulation, deterministic build-and-run evaluation design, experimental analysis, and paper writing, resulting in a **KDD 2026 ADS Track Main Conference** paper.

WORK EXPERIENCE

Tencent — Jan 2026 - Present
Project-Up Intern

- Served as technical owner for an internal innovation project, driving product roadmap, algorithm design, engineering implementation, data pipelines, benchmarks, post-training workflows, and experimental evaluation for production-oriented prototypes.

Microsoft Research Jan 2025 – Dec 2025

Research Intern, mentored by Zhongxin Guo

- Developed **AutoForge**, a framework that formulates agentic system design as a search problem over architectures, task decompositions, and tools, using Monte Carlo Tree Search to explore large spaces of agent variants across ARC-AGI, HumanEval, GSM8K, SWE-Bench, and ALFWorld.
- Designed a trajectory-guided hierarchical optimization loop over orchestration, decomposition, components, and parameters, with a decision optimizer that selects optimization targets based on execution traces.
- Extended the search space to cost-aware assignments of heterogeneous base LLMs, using capability-specific leaderboards and model pricing to preserve agent performance while reducing inference cost.

Microsoft Oct 2024 – Jan 2025

Software Engineer Intern

- Designed and implemented an internal front-end component library using TypeScript and Lit for Microsoft Edge, with Storybook documentation for cross-browser UI development.
- Developed Microsoft Edge Mobile features on Android and iOS, addressing responsive layout inconsistencies across devices to improve cross-platform UI reliability.

OPEN-SOURCE EXPERIENCE

ChromiumOS 21k stars

Google Summer of Code 2025 Contributor, mentored by Sarthak Kukreti

- Enhanced the farfetchd service by implementing tracing and replay support for disk I/O during app startup.
- Built a tracing API to log page-level disk accesses and a replay mechanism for preloading data.
- Validated performance gains through tast-based Go tests, measuring reductions in cold-start latency.

OpenManus-RL 4.1k stars

Project Leader

- Led development of the agent fine-tuning module of OpenManus, a reinforcement learning framework for fine-tuning LLM agents in ALFWorld, WebShop, and GAIA environments.
- Built online RL training infrastructure based on veRL and verl-agent; integrated benchmarks, generated rollout trajectories in ReAct, CoT, and modular formats, and built baseline comparisons with ragen.

libCacheSim | *A high-performance C++ library for cache simulation* 300 stars

Contributor

- Bridged high-performance C++ cache simulation with JavaScript by developing a Node.js binding.
- Authored and published the `libcachesim` package to the npm registry.

SKILLS

- **Languages**: Java, Python, C/C++, JavaScript, TypeScript, Go, SQL, Swift, Objective-C, Kotlin, HTML5/CSS3.
- **Frameworks & Libraries**: React, Node.js, Lit, Spring Boot, Spring Cloud, Django, Flask, Spring Data JPA.
- **Databases & Cloud Platforms**: MySQL, PostgreSQL, MongoDB, AWS, Google Cloud Platform.
- **Tools & Systems**: Git, Jenkins, CI/CD, Docker, Kubernetes, Kafka, RabbitMQ, Nginx, Prometheus, Grafana.

RELEVANT COURSES

Operating Systems, Database Systems, Distributed Systems, Algorithm Design and Analysis, Computer Networking, Software Engineering, Web Development, Compilers, Machine Learning, Deep Learning, Reinforcement Learning, Image Understanding, Natural Language Computing, Business of Software.

SERVICE

- ACL 2025 Reviewer
- ICML 2026 Reviewer