

Yimin Liu

Ph.D. Candidate, The Ohio State University (expected 2027)

yiminliu.career@gmail.com | 614-967-0122 | Columbus, OH

yiminl.com | Google Scholar | LinkedIn | GitHub

RESEARCH INTERESTS

I work on LLM agent infrastructure and evaluation — skill systems, harness design, and benchmarks (SkillsBench, ClawsBench) that measure whether agents actually perform across diverse tasks. I also develop deep learning methods for multi-omics integration in cancer biology.

SELECTED PUBLICATIONS

- **ClawsBench**: Li, X., Choe, K.W., **Liu, Y.**, Chen, X., et al. ClawsBench: Evaluating Capability and Safety of LLM Productivity Agents in Simulated Workspaces, *COLM 2026*
- **SkillsBench**: Li, X., **Liu, Y.**, Chen, W., Zheng, S., et al. SkillsBench: Benchmarking How Well Agent Skills Work Across Diverse Tasks, *Under review, NeurIPS 2026 D&B* (200+ citations)
- **BenchGuard**: BenchGuard: Model-Backed Instrumentation for Reward Integrity in LLM-Agent Evaluation Infrastructure, *Under review, USENIX Security 2027 (Cycle 1)*
- Zhang, C., **Liu, Y.**, Chen, X., Ji, P. What Keeps Agent Skills from Being Reusable? Evidence from 138K SKILL.md Files, *arXiv 2026*
- **DeepDR**: Jin, S., Jiang, Y., **Liu, Y.**, Ma, T., et al. DeepDR: An Integrated Deep-Learning Model Web Server for Drug Repositioning, *arXiv 2025*
- **DG-scRNA**: **Liu, Y.**, Gokbag, B., Majumdar, A., et al. Deep Learning with Graphic Cluster Visualization to Predict Cell Types of Single-Cell RNAseq Data, *ICIBM 2025, Oral Presentation*
- Liu, Y., Chen, Q., ..., **Liu, Y.**, et al. Proteomic profiling and genome-wide mapping of O-GlcNAc chromatin-associated proteins reveal an O-GlcNAc-regulated genotoxic stress response, *Nature Communications* 2020

RESEARCH EXPERIENCE

Research Scientist Intern & Open-Source Contributor

San Francisco, CA (Research Scientist Intern: May – Aug 2026)

BenchFlow

Jan 2026 – Present

- **Evaluation Infrastructure**: Built the BenchFlow evaluation environment and runtime powering execution of benchmark suites including SkillsBench, AgentFuzzBench, and FrontierPhysics (100+ containerized tasks) across coding, terminal, and productivity domains.
- **BenchGuard – Reward Integrity for LLM-Agent Evaluation (Under review, USENIX Security 2027)**:
 - Designed a model-backed instrumentation layer combining phase-aware static taint analysis with runtime infrastructure-side evidence to detect reward hacking in agent benchmarks.
 - Built *BenchGuard Trajectories*, a human-labeled corpus of 456 adjudicated trajectories from 31K+ public agent runs; achieved 96% detection accuracy and improved full-chain recall from 23–94% to 77–100%.
- **ClawsBench – LLM Agent Capability & Safety Evaluation (COLM 2026)**:
 - Built a benchmark that simulates 5 productivity services (Gmail, Slack, Google Calendar, Google Docs, Google Drive) with 44 tasks testing both capability and safety of LLM agents.
- **SkillsBench – LLM Agent Skill Benchmarking, 87 tasks / 8 domains / 18 model-harness configs (Under review, NeurIPS 2026)**:
 - Implemented the native skill system for Terminus2 so agents could load and run skills as modular packages; curated skills lift mean pass rate from 33.9% to 50.5% (+16.6 pp).
 - Developed the mHC layer task, a credentialed A100 GPU task implementing manifold-constrained hyper-connections in nanoGPT, bundling 3 reusable skills; designed ablation experiments to quantify per-skill contribution.
 - Built the analysis pipeline from raw trajectories to paper figures across all runs; led code review across the repo.
- **FrontierPhysics – Benchmarking AI Agents on Frontier Physics Research (ongoing)**: building the contribution review pipeline and task infrastructure (Docker environments, automated verifiers, rubric-based LLM-as-judge reviewer agents) for an open benchmark of PhD-level physics research tasks.

Independent Research Collaborator

Remote

AutoRSI & AgentFuzzBench

Jun 2026 – Present

Collaborators: Yuxuan Jiang (AutoRSI); Security Lab, UC Davis (AgentFuzzBench)

- **AutoRSI – Recursive Self-Improvement of Naive Methods (manuscript in preparation)**: a general framework that automatically strengthens a naive base method across domains via iterative propose–implement–evaluate–reflect cycles.
 - Building the RSI loop pipeline that proposes, implements, and evaluates candidate improvement plans, then reflects on failures and gains to drive the next iteration.

- Developing the core method: automatic construction and refinement of the improvement-plan pool, replacing fixed plan libraries that fail to generalize; extending to new domains to validate generalization.
- **AgentFuzzBench – Benchmarking End-to-End Agentic Fuzzing:** a benchmark evaluating agents that control the full fuzzing lifecycle (build, harnessing, fuzzing, crash triage) with repeated trials and ground-truth bugs.
 - Built the uniform task-package pipeline converting per-target specs into reproducible BenchFlow task packages (Docker environment, task contract, verifier) across 8 exact-pinned OSS targets.
 - Packaged the benchmark as an externally hosted BenchFlow benchmark with committed AFL++ oracle evidence; hardened reward integrity of the scored-coverage grading path.

Research Collaborator

Wuhan University of Science and Technology

Wuhan, CN (Remote)

Sep 2025 – Nov 2025

Collaborator: Shuting Jin, School of Computer Science and Technology

- **DeepDR – Drug Repositioning Web Platform:** Built the web platform: 15+ deep learning models over a 5.9M-edge knowledge graph, with visualizations explaining each recommendation.

Graduate Research Assistant

The Ohio State University

Columbus, OH

Mar 2023 – Present

Advisor: Lijun Cheng

- Developed DG-scRNA, a graph-convolutional network that annotates cell types from scRNA-seq data, 97.2% accuracy on 150K+ cells (first-author manuscript).
- Integrated scRNA-seq and spatial transcriptomics from 25 thyroid cancer samples; identified 50+ ligand-receptor interactions and two novel cell subtypes.
- Benchmarked CNN, RNN, and Transformer architectures for CRISPR gRNA on-target prediction over 50K+ curated sequences.

Research Assistant

Harvard Medical School

Boston, MA

Aug 2021 – Oct 2021

Advisor: Martin Hemberg

- Built automated mutation-calling pipeline (GATK, R) on 50K+ scRNA-seq cells; clustered SNV and expression profiles to predict variant–phenotype relationships in cancer.

Research Assistant

Beijing Institute of Genomics, Chinese Academy of Science

Beijing, CN

Mar 2021 – Aug 2021

Advisor: Mingkun Li

- Built a pipeline optimizing COVID-19 hybrid-capture probes; redesigned low-efficiency regions, improving sensitivity by 25%.

Research Assistant

Dalian University of Technology

Dalian, CN

Nov 2019 – Feb 2021

Advisor: Yubo Liu

- Undergraduate research on O-GlcNAc chromatin regulation in breast cancer (proteomics, ChIP-seq, RNA-seq); co-authored *Nature Communications* 2020, *Cell Biology International* 2021, and *BBA – General Subjects* 2021.

EDUCATION

The Ohio State University

Columbus, OH

Ph.D. Candidate in Biomedical Informatics, expected 2027 (Advisor: Lijun Cheng)

Aug 2022 – Present

Dalian University of Technology

Dalian, CN

B.S. in Bioinformatics, GPA: 86.4/100

Sep 2018 – Jun 2022

TECHNICAL SKILLS

LLM & Agent Evaluation: Benchmark Design, Eval Harnesses & Runtimes, Reward Integrity, LLM-as-Judge, Trajectory Scoring

Programming & ML: Python, PyTorch (GCNs, Transformers), NumPy/Pandas/scanpy, R (Seurat, Tidyverse), Bash, JavaScript

Bioinformatics & Genomics: scRNA-seq, Spatial Transcriptomics, Drug Repositioning, cellranger, samtools, STAR, bwa

Web, Cloud & Infrastructure: Full-Stack Development, REST APIs, Docker, HPC, Slurm, Git, Linux/Unix

ACADEMIC SERVICE

- **Workshop Organizer:** Agent Skills '26 – The First Workshop on Agent Skills, ACM CAIS 2026
- **Reviewer:** ICLR (2024–2026), ICML (2024–2025), NeurIPS (2024–2025), EMNLP (2026); PLOS Computational Biology