Nengneng Yu

+1 484-995-8987 | ynn1999@umd.edu | College Park, MD
samfisheryu.github.io | github.com/Samfisheryu

Personal Summary

My research builds **systems for LLMs**: observability and telemetry, fault-tolerant distributed training and serving, and collective communication. A complementary track in **data-driven ML** tackles real-world problems with messy, limited data.

Skills. C, C++, Python, PyTorch, NumPy, Pandas; LLM serving (vLLM, SGLang); distributed systems; GPU and collective communication; machine learning and deep learning.

Education

University of Maryland, College Park — College Park, MD
Ph.D. in Computer Science — GPA: 3.80/4.00 — *Aug 2023 – Present*
Advisor: Prof. Zaoxing (Alan) Liu

Boston University — Boston, MA
B.S. in Computer Engineering, Magna Cum Laude — *Sep 2019 – May 2023*

Research Projects

LLM & Distributed Systems

LLM Internal Observability — Sep 2025 – Present
Mentor: Prof. Zaoxing (Alan) Liu — *Froot Lab, UMD*

Inference-time workloads increasingly need timely access to a model's internal states, yet today's stacks treat observability as an afterthought; we pursue it as a first-class systems primitive.

- **DMI**: a high-performance, model- and engine-agnostic deep model inspector that decouples internal-state capture from the inference hot path *(arXiv preprint; poster at NSDI 2026).*

Reliable & Resilient Collective Communication for Distributed ML — May 2025 – Present
Mentor: Prof. Zaoxing (Alan) Liu | Collaborator: Wei Wang — *Froot Lab, UMD*

Network failures waste 10–15% of GPU hours in large-scale ML clusters, and today's NCCL-style CCLs crash on any error. We close the gap from both the systems and theory sides.

- **R2CC**: an NCCL-compatible CCL that keeps collectives running under inter-node NIC/link failures with negligible overhead *(arXiv preprint; full paper under review).*
- **OptCC**: the first information-theoretic lower bound on AllReduce time under bandwidth-asymmetric topologies, plus a pipelined algorithm that approaches it *(manuscript under review).*

Interactive Research Agents for Internet Incident Investigation — May 2023 – Nov 2023
Mentor: Prof. Zaoxing (Alan) Liu — *Froot Lab, UMD*

Investigating Internet incidents (outages, BGP misconfigurations, natural-disaster impacts) requires extensive cross-domain expertise and is largely manual today.

- **Generative Research Agent**: an LLM-based agent (Auto-GPT + GPT-4) that simulates an experienced researcher's investigation loop *(HotNets 2023; co-first author).*

Data-Driven ML

AI for Science — Cross-Cohort Biomedical Analysis Aug 2024 – Jun 2025
Mentors: Prof. Zaoxing (Alan) Liu, Yuefan Wang *Froot Lab, UMD & Johns Hopkins Med.*

Biomedical multi-omics is small-sample, high-dimensional, and batch-effect heavy across cohorts; we attack this with generative augmentation and integrative analysis.

- **TabSyM**: a generative pipeline (tabular diffusion + task-aware sampling + multi-domain adversarial alignment) that substantially boosts cancer-prognosis prediction across cohorts *(bioRxiv preprint; first author).*
- **15-layer multi-omics gastric cancer dissection**: a multi-institution study identifying therapeutically actionable gastric-cancer ecotypes *(Cell Reports Medicine; primary author).*

APT Detection under Concept Drift Feb 2022 – May 2023
Mentors: Prof. Zaoxing (Alan) Liu, Prof. Tuo Zhao *Red Hat & Boston Univ. & Georgia Tech*

ML-based provenance intrusion detection systems suffer concept drift as Advanced Persistent Threats evolve, while post-drift labels are scarce and naive retraining causes catastrophic forgetting.

- **Tidal**: a multi-head Transformer with a pre-train/fine-tune workflow that adapts to new attacks while retaining prior-attack accuracy *(NINeS 2026).*

Publications & Preprints

[1] **Nengneng Yu**, Sixian Xiong, Yibo Zhao, Wei Wang, Zaoxing Liu. "Enabling Performant and Flexible Model-Internal Observability for LLM Inference." *arXiv preprint*, 2026.

[2] **Nengneng Yu**, Sixian Xiong, Yibo Zhao, Wei Wang, Zaoxing Liu. "DMI: Performant and Flexible Deep Model Inspector for LLM Inference." *Poster, USENIX NSDI 2026.*

[3] Wei Wang, **Nengneng Yu**, Sixian Xiong, Zaoxing Liu. "Reliable and Resilient Collective Communication Library for LLM Training and Serving." *arXiv preprint*, 2025.

[4] **Nengneng Yu**, Yuefan Wang, Lindsey Kathleen Olsen, Bing Zhang, Hui Zhang, Zaoxing Liu. "TabSyM: A Generative Pipeline for Small Multi-Cohort Omics Tabular Data." *bioRxiv preprint*, 2025.

[5] Yuefan Wang*, Lindsey Kathleen Olsen*, . . . , **Nengneng Yu**, . . . , Bing Zhang. "A 15-Layer Multi-Omics Dissection of Gastric Cancer Ecotypes Reveals Therapeutic Opportunities." *Cell Reports Medicine*, 2026. *(**Nengneng Yu** as primary author.)*

[6] Yajie Zhou*, **Nengneng Yu***, Zaoxing Liu. "Towards Interactive Research Agents for Internet Incident Investigation." *HotNets 2023. (* co-first authors.)*

[7] Yajie Zhou, **Nengneng Yu**, Tuo Zhao, Zaoxing Liu. "Tidal: Tackling Concept Drift in Provenance-Based Advanced Persistent Threats Detection." *NINeS 2026.*