Nengneng Yu

+1 484-995-8987 | ynn1999@umd.edu | College Park, MD

Personal Summary

- **AI-driven systems researcher**: primary focus on **systems for LLMs**—observability/telemetry, distributed and fault-tolerant training/serving, collective communication and networking performance, with a complementary track in **data-driven ML**. I bridge algorithm design and systems engineering to deliver reliable, measurable, and scalable platforms.
- **Technical Proficiency:** C++/C, Python, PyTorch, Pandas, Numpy, LLM serving(vLLM,SGLang), Machine Learning, Deep Learning, system programming, data structures, and algorithms.

Education

University of Maryland College Park College Park, MD
Doctor of Philosophy in Computer Science, GPA: 3.80/4.00 *Aug 2023 - Present*
Advisor: Zaoxing(Alan) Liu ***https://zaoxing.github.io/***

Boston University Boston, MA
Bachelor of Science in Computer Engineering, Magna Cum Laude *Sep 2019 - May 2023*

Research Projects

LLM & Networking Systems

R2CC: Reliable Collective Communication for LLM Training & Serving May 2024 – Present
Mentor/Collaborator: Zaoxing(Alan) Liu/Wei Wang *Froot Lab, UMD*

- Enabled **seamless, lossless migration** and **failure-aware optimal scheduling** via **RDMA**, **multi-NIC GPU-memory preregistration**, and **NVLink/PCIe** topology.
- Evaluated across both **SimAI** and **2x8 H100 Cluster**, demonstrating robustness under injected failures for training and inference.
- Our **fault-tolerant collective communication library** outperforms recovery-based approaches (e.g., checkpoint/restart) and existing fault-tolerant frameworks in end-to-end **LLM training** and **serving** under failure happens. Full paper under review. Manuscript submitted to **Arxiv**.

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

- Developed an **LLM-based agent** to simulate experienced researchers and automate the investigation process, addressing the inefficiencies of traditional manual and time-consuming Internet incident investigations.
- Built an agent using **Auto-GPT** and **GPT-4**, equipped with autonomous **information retrieval, knowledge memory,** and **self-learning capabilities**. Tested it on challenging scenarios such as the impact of hypothetical solar storms on networks.
- Achieved 87.5% consistency in insights compared to human experts, effectively automating complex Internet incident analysis.
- Co-first authored paper appeared at **HotNet 2023**

Data-Driven ML

Generative AI for Cross-Cohort Biomedical Data Analysis Aug 2024 – Jun 2025
Mentor: Zaoxing(Alan) Liu, Yuefan Wang *Froot Lab, UMD & Johns Hopkins Medicine*

- Designed and built **TabSyM**: an end-to-end modular pipeline combining **tabular diffusion (TabDDPM)**, **task-aware sample selection**, and **multi-domain adversarial alignment (MDAN)** to address **small-sample, high-dimensional** omics and **batch effects**.
- Impact: **+30.2% AUROC** on gastric-cancer 3-year survival (five cohorts); up to **+22.1% AUROC / +21.8% F1** on pancreatic-cancer staging vs. State-Of-The-Art baselines.
- Led end-to-end research (design, protocols, ablations, interpretability, reproducibility); methodology manuscript submitted to **bioRxiv**; methods adopted in a collaborating **Cell** submission.

APT Detection and Analysis under Concept Drift Feb 2022 – May 2023
Mentor/Collaborator: Zaoxing(Alan) Liu, Tuo Zhao/Yajie Zhou *Red Hat & Boston University & Georgia Tech*

- Designed and developed **TIDAL**, a novel intrusion detection system to address **concept drift** in APT detection, where evolving attack patterns evade traditional ML-based defenses.
- Engineered a **Multi-head Transformer** architecture and a pre-train/fine-tune workflow to learn evolving attack patterns from limited data while preventing **catastrophic forgetting** of prior knowledge.
- Outperformed state-of-the-art systems in concept drift scenarios, achieving **27% higher recall** and **31% higher precision** on new attacks with 50% less training data, while retaining **43% higher recall** on previous attacks.

Publication & Works

- **[1]** Wei Wang, **Nengneng Yu**, Sixian Xiong, Zaoxing Liu, ”Reliable and Resilient Collective Communication Library for LLM Training and Serving”, Arxiv, 2025
- **[2]** **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, 2025
- **[3]** Yuefan Wang*, Lindsey Kathleen Olsen*, ..., **Nengneng Yu** (Primary Author),...,Bing Zhang, ”A 15-Layer Multi-Omics Dissection of Gastric Cancer Ecotypes Reveals Therapeutic Opportunities”, Under Cell review
- **[4]** Yajie Zhou*, **Nengneng Yu*** , Zaoxing Liu, ”Towards Interactive Simulacra of Internet Investigation by Human Researchers”, Hot Topics in Networks (HotNets), 2023
- **[5]** Yajie Zhou, **Nengneng Yu**, Chao Zhang, Tuo Zhao, Zaoxing Liu, ”Tackling Concept Drift in Provenance-based Advanced Persistent Threats Detection”, New Ideas in Networked Systems (NINeS), 2026

Miscellaneous Projects

Concurrency Control Schemes for Database Systems Jan 2023 – May 2024

- Implemented and evaluated six concurrency control schemes, including **Two-Phase Locking (2PL)**, **Optimistic Concurrency Control (OCC)**, and **Multi-Version Concurrency Control (MVCC)**. Developed two versions of 2PL with exclusive and shared locks, serial and parallel validation versions of OCC, and a simplified MVCC with **Serializable Snapshot Isolation (SSI)**.
- Developed and integrated the concurrency control schemes into a **main-memory key-value store** using **C++** and **thread management** techniques. Built a prototype transaction processing framework with a custom lock manager and multi-threaded execution support.
- Conducted performance benchmarking with **CMake** and **CTest** to assess throughput and latency across varying transaction lengths and contention levels.

eBPF Modularity Project Sep 2022 – Dec 2022

- Collaborated with professors from Brown University and IBM engineers to advance research on building a comprehensive and reusable **eBPF module library**.
- Leveraged **OPENED** tool to analyze and decompose eBPF programs from open-source projects, extracting reusable modules for improved modularity and maintainability.
- Designed a framework to transform extracted modules into a format compatible with **Bumblebee** tools for generating OCI images. Integrated “glue logic” for seamless module compatibility with **L3AF** and **Polycube**.

Basic Unix-like Operating System Jan 2022 – May 2022

- Implemented a custom shell capable of executing commands via system calls, redirecting stdin/stdout, handling multiple command pipelines, and supporting background execution with “&”.
- Developed a subset of the **POSIX threads API** in user mode, enabling multi-threaded execution with round-robin scheduling for effective resource management.
- Designed and implemented a copy-on-write (COW) thread-local storage (TLS) mechanism to enable data sharing between threads while ensuring isolation of changes, improving memory management at the thread level.