

Yuanli Wang

(+1) 612-963-5189 | yuanliw@bu.edu | <https://pentium3.github.io/>

EDUCATION

Boston University

Boston, MA

PhD in Computer Science

09/2021 – 05/2027 (*Expected*)

- Advisor: Prof. Vasiliki Kalavri

University of Minnesota, Twin Cities

Minneapolis, MN

Master of Science in Computer Science

09/2018 – 06/2021

RESEARCH EXPERIENCE

Deadline-Aware LLM Agent Scheduling on Hybrid Cloud

07/2026 – *Present*

- Developing a training-free scheduling framework that routes concurrent LLM agent requests between local and cloud inference to meet per-task deadlines while minimizing cloud cost.
- Extending the scheduling policy to account for KV-cache rebuilding overhead during local-cloud switching.

Benchmarking LLM Agents for Research Artifact Deployment

01/2026 – 07/2026

- Designed DeployBench, a benchmark for evaluating LLM agents on end-to-end research artifact deployment, covering 51 tasks across AI/ML, computer systems, and scientific computing.
- Built a realistic evaluation pipeline with paper-specific hidden verifiers, requiring agents to install dependencies, repair toolchain issues, configure GPU/CUDA or system-level components, and execute designated paper experiments.
- Evaluated 4 frontier LLM agents using OpenHands, with the strongest achieving only a 51% pass rate. Analyzed failed trajectories and identified premature agent self-termination as a dominant failure mode.

Evaluating Capability and Safety of LLM Productivity Agents

01/2026 – 06/2026

- Contributed to ClawsBench, a benchmark evaluating LLM productivity agents across 5 high-fidelity simulated workspace environments and 44 tasks, covering cross-service workflows and safety-critical scenarios.
- Conducted evaluation experiments across multiple LLMs and agent harnesses, analyzing the impact of agent scaffolding on task success and identifying unsafe behaviors such as confidential data leakage and unauthorized contract modification.

Graph-Based Trajectory Evaluation for Web Agents

05/2025 – 10/2025

- Designed a graph-based evaluation framework that transforms web-agent trajectories into per-task weighted action graphs through LLM-based annotation, canonicalization, and deterministic graph merging.
- Validated the framework on 4,768 trajectories across 812 tasks, developing graph-based reward metrics to identify inefficient actions and critical steps overlooked by outcome-only evaluation.

PUBLICATIONS

- [1]. DeployBench: Benchmarking LLM Agents for Research Artifact Deployment. **Yuanli Wang**, Yaoyao Qian, Yue Zhang, Hanhan Zhou, Jindan Huang, Tianfu Fu, Qiuyang Mang, Huanzhi Mao, Wenhao Chai, Wendong Fan, Liqiang Jing. EMNLP 2026 Findings.
- [2]. ClawsBench: Evaluating capability and safety of LLM productivity agents in simulated workspaces. Xiangyi Li, Kyoung Whan Choe, Yimin Liu, Xiaokun Chen, Chujun Tao, Bingran You, Wenbo Chen, Zonglin Di, Jiankai Sun, Shenghan Zheng, Jiajun Bao, **Yuanli Wang**, Weixiang Yan, Yiyuan Li, Han-chung Lee. COLM 2026.
- [3]. Terminal-Bench: Benchmarking Agents on Hard, Realistic Tasks in Command Line Interfaces. Terminal-Bench Team (including **Yuanli Wang**). ICLR 2026.

- [4]. SkillsBench: Benchmarking how well agent skills work across diverse tasks. SkillsBench Team (including **Yuanli Wang**). NeurIPS 2026.
- [5]. WebGraphEval: Multi-Turn Trajectory Evaluation for Web Agents using Graph Representation. Yaoyao Qian, **Yuanli Wang**, Jinda Zhang, Yun Zong, Meixu Chen, Hanhan Zhou, Jindan Huang, Yifan Zeng, Xinyu Hu, Chan Hee Song, Danqing Zhang. NeurIPS 2025 Workshop on Multi-Turn Interactions in Large Language Models.
- [6]. CAPSys: Contention-aware task placement for data stream processing. **Yuanli Wang**, Lei Huang, Zikun Wang, Vasiliki Kalavri, and Abraham Matta. EuroSys 2025.
- [7]. The *Non-Expert Tax*: Quantifying the cost of auto-scaling in Cloud-based data stream analytics. **Yuanli Wang**, Baiqing Lyu, Vasiliki Kalavri. BiDEDE 2022.
- [8]. A New Benchmark Harness for Systematic and Robust Evaluation of Streaming State Stores. Esmail Asyabi, **Yuanli Wang**, John Liagouris, Vasiliki Kalavri, Azer Bestavros. EuroSys 2022.
- [9]. HACCS: Heterogeneity-Aware Clustered Client Selection for Accelerated Federated Learning. Joel Wolfrath, Nikhil Sreekumar, Dhruv Kumar, **Yuanli Wang**, Abhishek Chandra. IPDPS 2022.
- [10]. Accelerated Training via Device Similarity in Federated Learning. **Yuanli Wang**, Joel Wolfrath, Nikhil Sreekumar, Dhruv Kumar, Abhishek Chandra. EdgeSys 2021.
- [11]. Fail-slow fault tolerance needs programming support. Andrew Yoo, **Yuanli Wang**, Ritesh Sinha, Shuai Mu, Tianyin Xu. HotOS 2021.

PROFESSIONAL EXPERIENCE

Megagon Labs | *Research Intern*

05/2025 – 08/2025

- Designed a multi-agent system for decision-making under uncertainty, which uses LLMs and probabilistic inference to automatically discover and model causal relationships between hidden factors from data.

Apple | *Data Processing Platform Intern*

05/2023 – 08/2023

- Integrated OpenLineage framework with Flink to track data lineage for Apple’s AIML data processing platform. Implemented new OpenLineage visitors for Kafka and Iceberg data connectors in Flink.
- Contributed to OpenLineage open source repository.

Google Summer of Code (CNCF / PingCAP) | *Software Engineering Intern*

05/2019 – 08/2019

- Worked on AutoTiKV project from scratch: used machine learning to tune a database system under user-specific workloads.
- Implemented a Gaussian Process Regression Model to predict the performance of RocksDB (the core storage engine in TiKV) under different knob configurations. Achieved 1.3x lower latency under several types of workloads without human guidance.

PROFESSIONAL SERVICE

- **Reviewer:** ICLR 2027, NeurIPS 2026, ICLR 2026, ICASSP 2026, ICLR 2025 , ICASSP 2025

SKILLS

Programming: Python, Java, C++, Bash, SQL

LLM Agents & Inference: OpenHands, Harbor, vLLM, SGLang

Systems & Infrastructure: Linux, Docker, Git, Flink, Kafka, AWS, GCP

SELECTED AWARDS

Rank 16/183 in 2018 ACM-ICPC North Central North America Regional Contest	11/2018
Bronze Medal, China Collegiate Programming Contest	10/2015