Yufan (Ethan) Wu

Ph.D. Student, Computer Science and Engineering, The Ohio State University

wu.6545@osu.edu | umwyf.github.io/my-website | Google Scholar

Education

The Ohio State University — Columbus, OH
Ph.D. in Computer Science and Engineering (AI Track) — *Sep. 2024 – Aug. 2029 (Expected)*

University of Michigan — Ann Arbor, MI
B.S. in Computer Science and Engineering — *Sep. 2022 – Aug. 2024*

Shanghai Jiao Tong University — Shanghai, China
B.S. in Electrical and Computer Engineering (UM–SJTU Joint Institute) — *Sep. 2020 – Aug. 2024*

Research Interests

LLM reasoning & agents: weak-to-strong generalization, search and deep-research agents, tool-using agents, agent memory, theory-of-mind evaluation.

Efficient AI systems: data-efficient Earth observation and onboard intelligence for LEO satellite networks.

Selected Papers

* Equal contribution. Full list in All Papers.

[1] Hi-ToM: A Benchmark for Evaluating Higher-Order Theory of Mind Reasoning in Large Language Models
Yufan Wu*, Yinghui He*, Yilin Jia, Rada Mihalcea, Yulong Chen, Naihao Deng
Findings of EMNLP, 2023 – 100+ citations

- Released a benchmark testing higher-order theory-of-mind abilities of LLMs; built an automated data generator and coordinated human annotation of 600+ samples.

[2] CritICL: Inference-Time Weak-to-Strong Generalization from Small Language Model Failure Modes
Yufan Wu, Yinghui He, Zhengyi Hu, Lang Wei, Ruichen Li, Qifan Yang, Ting Zhu
Conference on Language Modeling (COLM), 2026

- Proposed an inference-time framework that improves LLM reasoning by retrieving critiques derived from small-model failure modes.
- Built the failure-mode retrieval and critique-generation pipeline; achieved higher accuracy at lower inference cost across multiple reasoning benchmarks.

[3] QUEST: Training Frontier Deep Research Agents with Fully Synthetic Tasks
Jian Xie*, Tianhe Lin*, Zilu Wang*, . . . , **Yufan Wu**, et al.
Conference on Neural Information Processing Systems (NeurIPS), 2026

- A family of 2B–35B open deep-research agents trained entirely on synthetic research tasks, spanning fact seeking and long-form report synthesis.

[4] FLAP: Failure-Aware Planning for Search Agents
Yufan Wu, Yinghui He, Zhengyi Hu, Lang Wei, Ruichen Li, Qifan Yang, Ting Zhu
Under review at NAACL 2027

- Separated search planning from execution with evidence verification and local recovery.
- Built the planner training pipeline; outperformed RL-based search agents by up to 8.5% across seven QA benchmarks while improving retrieval quality and failure recovery.

Research Experience

Research Assistant, The Ohio State University — Sep. 2024 – Present
Advisor: Prof. Ting Zhu — *Columbus, OH*

Research Assistant, LIT Group, University of Michigan — Jan. 2023 – May 2024
Advisor: Prof. Rada Mihalcea — *Ann Arbor, MI*

Research Assistant, University of Michigan — Sep. 2022 – Jan. 2023
Advisor: Prof. Michael P. Wellman — *Ann Arbor, MI*

All Papers

* Equal contribution.

Conference Papers

[C1] QUEST: Training Frontier Deep Research Agents with Fully Synthetic Tasks
Jian Xie*, Tianhe Lin*, Zilu Wang*, . . . , **Yufan Wu**, et al.
Conference on Neural Information Processing Systems (NeurIPS), 2026

[C2] CritICL: Inference-Time Weak-to-Strong Generalization from Small Language Model Failure Modes
Yufan Wu, Yinghui He, Zhengyi Hu, Lang Wei, Ruichen Li, Qifan Yang, Ting Zhu
Conference on Language Modeling (COLM), 2026

[C3] Achieving Efficient Storage and Communication via Collaboration
Ruichen Li*, **Yufan Wu***, Zhengyi Hu*, Sheng-Jyun Cai, Lang Wei, Qifan Yang, Ting Zhu
ACM SIGCOMM, 2026 (acceptance rate 16.2%)

[C4] Exploring LEO Satellite Networks for Continuous Monitoring and Dynamic Tracking
Lang Wei, Ruichen Li, Baodong Chen, Qifan Yang, **Yufan Wu**, Ting Zhu
ACM Conference on Embedded Networked Sensor Systems (SenSys), 2026 (acceptance rate 19.8%)

[C5] Hi-ToM: A Benchmark for Evaluating Higher-Order Theory of Mind Reasoning in Large Language Models
Yufan Wu*, Yinghui He*, Yilin Jia, Rada Mihalcea, Yulong Chen, Naihao Deng
Findings of EMNLP, 2023 – 100+ citations

Under Review & Manuscripts

[U1] MemPO: Optimizing Memory Write Policies with Retrieval-Attributed Credits
Yufan Wu, Zhengyi Hu, Lang Wei, Ruichen Li, Qifan Yang, Laixi Shi, Ting Zhu
Under review at ICLR 2027

[U2] FLAP: Failure-Aware Planning for Search Agents
Yufan Wu, Yinghui He, Zhengyi Hu, Lang Wei, Ruichen Li, Qifan Yang, Ting Zhu
Under review at NAACL 2027

[U3] Just-in-Time Satellite Data Acquisition without Compression in LEO Satellite Networks
Yufan Wu*, Ruichen Li*, Zhengyi Hu*, Lang Wei, Qifan Yang, Ting Zhu
Under review at ACM SIGCOMM 2026

[U4] SafeTap: A Benchmark on Unsafe Transitions in Mobile Task Automation
Ruichen Li*, Sheng-Jyun Cai, Lang Wei, Qifan Yang, **Yufan Wu**, Ting Zhu
Under review at ACM SIGCOMM 2026

[U5] Entropy-Driven Scanning Optimization for Near Real-Time Earth Observation
Lang Wei, Ruichen Li, **Yufan Wu**, Qifan Yang, Ting Zhu
Under review at ICML 2026

[U6] **PAMELA: Probabilistic Amplification Module for Lightweight AI on LEO Satellites**
Yufan Wu*, Qifan Yang*, Lang Wei, Ruichen Li, Ting Zhu
Manuscript

Teaching & Advising

Teaching Assistant, Computer Networking Sep. 2024 – Aug. 2025
The Ohio State University

Teaching Assistant, Physics II Sep. 2021 – Aug. 2022
Shanghai Jiao Tong University

Student Advisor, UM–SJTU Joint Institute Advising Center Sep. 2021 – Aug. 2022
Shanghai Jiao Tong University

Academic Service

Reviewer, ACM SenSys 2025 Oct. 2024
Reviewer, ACL 2024 Aug. 2023

Honors & Awards

Dean's List (Outstanding GPA), University of Michigan Jan. 2023
Finalist Award (top 2%), Mathematical Contest in Modeling, COMAP May 2022
Undergraduate Merit Scholarship (top 5%), Shanghai Jiao Tong University Oct. 2021

Skills

Programming:	Python, C++, R
ML Frameworks:	PyTorch
LLM Training:	Reinforcement learning for LLM agents, training on synthetic tasks, planner training
LLM Inference:	Retrieval-augmented generation, in-context learning, critique generation, weak-to-strong generalization
LLM Agents:	Search and deep-research agents, agent memory, planning and failure recovery
Evaluation & Data:	Benchmark design, automated data generation, human annotation coordination, evaluation across reasoning and QA benchmarks
Systems & Networking:	LEO satellite networks, Earth-observation data pipelines, satellite scheduling, lightweight onboard models