

Yuheng Wu

Curriculum Vitae

☎ 650-833-8588
🌐 joel-wu.github.io
✉ yuhengwu@stanford.edu

Research Interests

Reasoning in large language models (LLMs), efficient test-time methods and scaling laws, LLMs for programming, verification and formal methods, machine learning systems, and hardware-software co-design.

Education

09.2026 – Now **Stanford University**, Stanford, California, USA.
Ph.D. in Electrical Engineering

09.2024 – 06.2026 **Stanford University**, Stanford, California, USA.
Master of Science in Electrical Engineering

09.2020 – 06.2024 **Wuhan University**, Wuhan, Hubei, China.
Bachelor of Engineering in Electronic Information Engineering

Publications

Preprint

- Preprint Yuzong Chen*, Xilai Dai*, Jake Hyun*, Chi-Chih Chang, Wonsuk Jang, **Yuheng Wu**, Thierry Tambe, Jae-sun Seo, and Mohamed S. Abdelfattah. “RaZeR: Pushing the Limits of NVFP4 Quantization with Redundant Zero Remapping,” *Arxiv Preprint*, Jan. 2026
<https://arxiv.org/abs/2501.04052>
- Preprint **Yuheng Wu**, Berk Gokmen, Zhouhua Xie, Peijing Li, Caroline Trippel, Priyanka Raina, and Thierry Tambe. “LLM-FSM: Scaling LLMs for Finite-State Reasoning in RTL Code Generation,” *Arxiv Preprint*, Feb. 2026
<https://arxiv.org/abs/2602.07032>
- Preprint **Yuheng Wu**, Azalia Mirhoseini, and Thierry Tambe. “On the Role of Temperature Sampling in Test-Time Scaling,” *Arxiv Preprint*, Oct. 2025
<https://arxiv.org/abs/2510.02611>

Conference

- ISCA’26 Yuzong Chen, Chao Fang, Xilai Dai, **Yuheng Wu**, Thierry Tambe, Marian Verhelst, and Mohamed S. Abdelfattah. “ P^3 -LLM: Unlocking Efficient Processing-In-Memory for Edge LLM Inference with Hybrid Numerical Formats,” in *Proceedings of the Annual International Symposium on Computer Architecture*, Jun. 2026
<https://arxiv.org/abs/2511.06838>
- MLSys’26 Hojoon Kim, **Yuheng Wu**, and Thierry Tambe. “AgenticCache: Cache-Driven Asynchronous Planning for Embodied AI Agents,” in *Proceedings of the Conference on Machine Learning and Systems*, May. 2026
<https://arxiv.org/abs/2604.24039>

- EMNLP'25 **Yuheng Wu**, Jianwen Xie, Denghui Zhang, and Zhaozhuo Xu. "DEL-ToM: Inference-Time Scaling for Theory-of-Mind Reasoning via Dynamic Epistemic Logic," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Nov. 2025.
<https://aclanthology.org/2025.emnlp-main.573>
- EMNLP'25 Anjiang Wei*, **Yuheng Wu***, Yingjia Wan, Tarun Suresh, Huanmi Tan, Zhanke Zhou, Sanmi Koyejo, Ke Wang, and Alex Aiken. "SATBench: Benchmarking LLMs' Logical Reasoning via Automated Puzzle Generation from SAT Formulas," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Nov. 2025.
<https://aclanthology.org/2025.emnlp-main.1716>
- COLM'25 Anjiang Wei, Tarun Suresh, Jiannan Cao, Naveen Kannan, **Yuheng Wu**, Kai Yan, Thiago S. F. X. Teixeira, Ke Wang, and Alex Aiken. "CodeARC: Benchmarking Reasoning Capabilities of LLM Agents for Inductive Program Synthesis," in *Proceedings of the Conference on Language Modeling*, Oct. 2025.
<https://openreview.net/forum?id=Q5pVZCrrKr>

Journal

- NPJ AI **Yuheng Wu**, Wentao Guo, Zirui Liu, Heng Ji, Zhaozhuo Xu, and Denghui Zhang. "How Large Language Models Encode Theory-of-Mind: A Study on Sparse Parameter Patterns," *Nature Partner Journals on Artificial Intelligence*, 1, 20, 2025.
<https://www.nature.com/articles/s44387-025-00031-9>

Honors and Awards

- 2025 Lambda Research Grant
- 2024 Outstanding Graduate Scholarship, Wuhan University
- 2021, 2023 National Scholarship, Ministry of Education, China
- 2021, 2022, 2023 Merit Student, Wuhan University
- 2023 Outstanding Young Volunteer, Wuhan University
- 2022 Yu Gang - Song Xiao Scholarship, Wuhan University

Service

- Reviewer ICLR, NeurIPS, ECCV, IEEE Transactions on Image Processing, Journal of Artificial Intelligence Research, IEEE Transactions on Neural Networks and Learning Systems, Pattern Recognition

Presentations

- Invited Talk Sparsity and Scaling: Unveiling and Enhancing Theory-of-Mind in Large Language Models, Advances in Sequence modeling from Algorithmic Perspectives (ASAP) Seminar, Online, Dec. 2025
- Invited Talk Can Large Language Models Solve SAT Problems, Stanford Microwave Integrated Circuits (SMIRc) Laboratory, Stanford University, Oct. 2025
- Poster On the Role of Temperature Sampling in Test-Time Scaling, Stanford Center for Portable Accelerated Learning (PORTAL) Annual Retreat, Half Moon Bay, Aug. 2025