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

09.2024 – now **Stanford University**, Stanford, California, USA.
Master of Science in Electrical Engineering, GPA: 4.05

09.2020 – 06.2024 **Wuhan University**, Wuhan, Hubei, P.R.China.
Bachelor of Engineering in Electronic Information Engineering, GPA: 3.98

Publications

Preprint

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," *Preprint*, Nov. 2025
`https://joel-wu.github.io/pdf/LLM-FSM.pdf`

Preprint Yuzong Chen, Chao Fang, Xilai Dai, **Yuheng Wu**, Thierry Tambe, Marian Verhelst, and Mohamed Abdelfattah. "Unlocking Efficient Processing-In-Memory for Edge LLM Inference with Hybrid Numerical Formats," *Arxiv Preprint*, Nov. 2025
`https://arxiv.org/abs/2511.06838`

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

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`

CVPR'24 Yang Yu*, Erting Pan*, Xinya Wang, **Yuheng Wu**, Xiaoguang Mei, and Jiayi Ma. "Unmixing before Fusion: A Generalized Paradigm for Multi-Source-Based Hyperspectral Image Synthesis," *in Proceedings of the Conference on Computer Vision and Pattern Recognition*, Jun. 2024.
`https://ieeexplore.ieee.org/document/10656158`

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`

Workshop

NeurIPS'25 **Yuheng Wu** and Thierry Tambe. "On the Role of Temperature Sampling in Test-Time Scaling," *in Efficient Reasoning Workshop and Foundations of Reasoning in LMs Workshop at the Conference on Neural Information Processing Systems*, Dec. 2025

ICML'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 Efficient Systems for Foundation Models Workshop at the International Conference on Machine Learning*, Jul. 2025

ICLR'25 **Yuheng Wu**, Wentao Guo, Zirui Liu, Zhaozhuo Xu, and Denghui Zhang. "Sensitivity Meets Sparsity: The Impact of Extremely Sparse Parameter Patterns on Theory-of-Mind of Open-Source Large Language Models," *in Open Science for Foundation Models Workshop at the International Conference on Learning Representations*, Apr. 2025.

Experience

09.2025 – now Research Intern, LLMs for RTL Code Generation, Stanford University
Mentor/Collaborator: **Thierry Tambe**, **Priyanka Raina**, and **Caroline Trippel**
• Built a scalable spec-to-RTL data generation pipeline using LLMs and SAT-based formal verification, providing controlled finite-state and temporal reasoning tasks for RTL synthesis and evaluation. [Preprint]

05.2025 – now Research Intern, Test-Time Methods for LLMs, Stanford University
Mentor/Collaborator: **Thierry Tambe** and **Azalia Mirhoseini**
• Revealed the saturation limit of sample-based test-time scaling and introduced temperature scaling as a new axis to unlock LLMs' reasoning potential, with an efficient multi-temperature voting method to reduce inference overhead. [Preprint]

03.2025 – 09.2025 Research Intern, Logical and Inductive Reasoning, Stanford University
Mentor/Collaborator: **Alex Aiken** and **Sanmi Koyejo**
• Built a SAT solver grounded framework for natural language logical puzzle generation and evaluation with consistency checking, revealing systematic failures of LLMs in search based logical reasoning. [EMNLP'25]
• Assisted in developing a training data generation pipeline for inductive reasoning using GPT models, supporting work on inductive program synthesis through differential testing feedback. [COLM'25]

06.2024 – 09.2025 Research Intern, Reasoning and Interpretability, Stevens Institute of Tech. & UIUC
Mentor/Collaborator: **Zhaozhuo Xu**, **Denghui Zhang**, and **Heng Ji**
- Developed a formal-method-based framework for inference-time scaling of theory-of-mind reasoning in LLMs, grounded in dynamic epistemic logic; enabled verifiable belief-trace reasoning via a learned process belief model that supervises step-level belief updates during inference. [EMNLP'25]
- Identified Fisher-informative sparse parameters in LLMs that modulate query-key angles in attention heads, revealed RoPE-related effects, and linked these mechanisms to downstream theory-of-mind reasoning capabilities. [NPJ AI]

06.2023 – 05.2024 Research Intern, Multi-Spectral Computational Imaging, Wuhan University
Mentor/Collaborator: **Xiaoguang Mei** and **Jiayi Ma**
- Assisted in developing an asymmetric autoencoder with physical constraints in the latent space, integrating RGB inputs with diffusion and normalizing-flow models to synthesize credible and diverse hyperspectral data. [CVPR'24]

Honors and Awards

2024 Outstanding Graduate Scholarship, Wuhan University

2021, 2023 National Scholarship, Ministry of Education, P.R.China

2021, 2022, 2023 Merit Student, Wuhan University

2023 Outstanding Young Volunteer, Wuhan University

2022 Yu Gang - Song Xiao Scholarship, Wuhan University

Service

Reviewer CVPR 2026, ICLR 2026, NeurIPS 2025, IEEE Transactions on Image Processing, Journal of Artificial Intelligence Research, IEEE Transactions on Neural Networks and Learning Systems, Pattern Recognition

Grader EE 263: Matrix Methods, Stanford University, 2025 Fall

Presentations

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