

Xiangchen Song

☎ +1 412 880 3932 | ✉ xiangchs@cs.cmu.edu | 🌐 xiangchensong.github.io | 📄 Google Scholar

RESEARCH INTERESTS

LLM Alignment & Safety · Interpretability · Representation Learning · Agents & Evaluation

I aim to understand and shape the behavior of increasingly capable AI systems so that they remain aligned, reliable, and safe in deployment. One question runs through my work: *when can we trust what we recover from a model's internals — and what we claim about its behavior?*

- **Trustworthy interpretability** – Cross-run consistent [1] and provably identifiable [2] sparse autoencoders, grounded in my representation learning theory [3], [4], for reliable and reproducible interpretation.
- **Monitoring & oversight from internals** – Jacobian-lens analysis of frontier-scale open-weight models [5], and hidden-state probes for verifying reasoning steps and accelerating inference in large reasoning models [6].
- **Evaluation & post-training** – Testing whether test-time training installs reusable memory [7]; evaluating causal reasoning in LLM agents [8]; developing process supervision for diffusion LLMs [9]; and benchmarking long-horizon agent workflows [10].

EDUCATION

Carnegie Mellon University

Aug. 2021 – Dec. 2026 (expected)

Ph.D. in Machine Learning

M.S. in Machine Learning Research

May 2023

Advisor: [Kun Zhang](#)

University of Illinois Urbana-Champaign

Dec. 2020

B.S. in Mathematics and Computer Science

Summa Cum Laude, Bronze Tablet

Advisor: [Jiawei Han](#)

INDUSTRY EXPERIENCE

Amazon Applied Scientist Intern – Efficient Reasoning

May 2025 – Aug. 2025

- Developed hidden-state probes that verify reasoning steps and accelerate inference in large reasoning models.

Amazon Applied Scientist Intern – Speculative Decoding

May 2024 – Aug. 2024

- Proposed lossy speculative decoding methods that improve throughput by 10% with negligible quality loss.

Salesforce Research Research Intern – Time Series Representation

May 2022 – Nov. 2022

- Designed unsupervised temporal representation learning methods with identifiability guarantees.

Microsoft Research Asia Research Intern – Graph Mining

Mar. 2021 – Aug. 2021

- Built taxonomy-aware graph neural networks for personalized session-based recommendation.

SELECTED PUBLICATIONS & TECHNICAL REPORTS

Complete List: <https://xiangchensong.github.io/publications> (* indicates equal contribution)

Trustworthy Interpretability

- [1] Mechanistic Interpretability Should Prioritize Feature Consistency in Sparse Autoencoders
Xiangchen Song*, Aashiq Muhamed*, Yujia Zheng, Lingjing Kong, Zeyu Tang, Mona T. Diab, Virginia Smith, and Kun Zhang
ACL 2026 (Oral) [GitHub], Earlier version: *MechInterp Workshop NeurIPS 2025 (Spotlight)*
- [2] LLM Interpretability with Identifiable Temporal–Instantaneous Representation
Xiangchen Song*, Jiaqi Sun*, Zijian Li, Yujia Zheng, and Kun Zhang
NeurIPS 2025 [GitHub]
- [3] Causal Temporal Representation Learning with Nonstationary Sparse Transition
Xiangchen Song, Zijian Li, Guangyi Chen, Yujia Zheng, Yewen Fan, Xinshuai Dong, and Kun Zhang
NeurIPS 2024 [GitHub]

- [4] Temporally Disentangled Representation Learning under Unknown Nonstationarity
Xiangchen Song, Weiran Yao, Yewen Fan, Xinshuai Dong, Guangyi Chen, Juan Carlos Niebles, Eric Xing, and Kun Zhang
NeurIPS 2023 [GitHub]
-
- Monitoring & Oversight from Internals**
- [5] Testing the global workspace at open frontier scale: the Jacobian lens on DeepSeek-V4-Flash and GLM-5.2
Xiangchen Song, Fan Feng and Yujia Zheng
Research blog, Aug. 2026 [Blog] [GitHub]
- [6] Accelerating Speculative Reasoning with Internal Probing
Xiangchen Song, Saket Dingliwal, Sai Muralidhar Jayanthi, Vivek Govindan, Khiem Pham, Shobha Vasudevan, Beidi Chen, Sravan Babu Bodapati, and Aram Galstyan
Efficient Reasoning Workshop NeurIPS 2025
-
- Evaluation & Post-Training**
- [7] Beyond Perplexity: A Behavioral Evaluation Framework for Deployment-Memory Claims in LLM Test-Time Training
Xiangchen Song, Zhenhao Chen, Lingjing Kong, Shaoan Xie, Xinshuai Dong, Guangyi Chen, Kun Zhang
arXiv 2026
- [8] CausalGame: Benchmarking Causal Thinking of LLM Agents in Games
Zhenhao Chen, Yongqiang Chen, Chenxi Liu, Junchi Yu, **Xiangchen Song**, Zijian Li, Jialin Li, Philip Torr, Bo Han, and Kun Zhang
ICML 2026 (Oral)
- [9] Advancing Reasoning in Diffusion Language Models with Denoising Process Rewards
Shaoan Xie*, Lingjing Kong*, **Xiangchen Song**, Xinshuai Dong, Guangyi Chen, Eric P. Xing, and Kun Zhang
ACL 2026
- [10] χ -Bench: Can AI Agents Automate End-to-End, Long-Horizon, Policy-Rich Healthcare Workflows?
Haolin Chen, Deon Metelski, Leon Qi, Tao Xia, Joonyul Lee, Steve Brown, Kevin Riley, Frank Wang, T. Y. Alvin Liu, Hank Capps MD, Zeyu Tang, **Xiangchen Song**, Lingjing Kong, Fan Feng, Tianyi Zeng, Zhiwei Liu, Zixian Ma, Hang Jiang, Fangli Geng, Yuan Yuan, Chenyu You, Qingsong Wen, Hua Wei, Yanjie Fu, Yue Zhao, Carl Yang, Biwei Huang, Kun Zhang, Caiming Xiong, Sanmi Koyejo, Eric P. Xing, Philip S. Yu, and Weiran Yao
arXiv 2026

SERVICE

Organizer: ICDM 2024 Workshop on Causal Representation Learning.

Reviewer: NeurIPS, ICML, ICLR, ARR, AAI, UAI, CLear, AISTATS, KDD, CVPR, ICCV/ECCV.

HONORS & AWARDS

Thinking Machines Lab Tinker Research Grant (2026)

ICML Gold Reviewer (2026)

Modal Academic Research Grant (2026)

CMU MLD Presidential Fellowship (2025)

NeurIPS Scholar Award (2025)

CRA Outstanding Undergraduate Researcher Award (Honorable Mention, 2021)

UIUC John R. Pasta Outstanding Undergraduate Award (2020)

CityU HK Compass Scholarships (2018)

TECHNICAL SKILLS

Programming & ML: Python, PyTorch, Hugging Face Transformers, vLLM, SGLang, Slurm, Git.