

YUANPU CAO

Website: caoyuanpu.github.io

Tel: +1 646-209-5783

Email: ymc5533@psu.edu

EDUCATION

The Pennsylvania State University

College of Information Sciences and Technology

Ph.D. in Information Sciences and Technology, GPA: 3.83/4.0

- Advisor: Prof. Jinghui Chen
- Herbert J. Bailey Graduate Fellowship, 2026–2027 (*college's sole nominee*)

State College, PA, USA

Aug. 2023- May 2028 (Expected)

New York University

Department of Computer Science, Courant Institute of Mathematical Sciences

Master of Science in Computer Science, GPA: 3.75/4.0

New York, NY, USA

Aug. 2021- May 2023

Nankai University

College of Software

Bachelor of Engineering in Software Engineering, GPA: 3.84/4.0

Tianjin, China

Sep. 2016 - Jul. 2020

RESEARCH INTERESTS

My research focuses on **language models**, particularly **post-training for reasoning and agentic capabilities**, as well as **model steering and reliability** from a **hidden representation** perspective.

INTERNSHIP EXPERIENCE

Bloomberg - AI Research Intern

Jun. 2026 - Sep. 2026

- Developed a knowledge graph-guided reinforcement learning method that verifies evidence-retrieval steps against the reasoning topology and assigns process rewards without LLM-based judges, improving evidence-grounded multi-hop reasoning in 1.7B–8B language models.
- Built an iterative self-improvement pipeline using knowledge graph-guided, self-generated reasoning trajectories for fine-tuning before reinforcement learning, improving counterfactual robustness across three QA benchmarks without an external teacher and requiring no knowledge graph at inference time.

Amazon Web Services - Applied Scientist Intern, Agentic AI

May 2025 - Aug. 2025

- Developed a summarization-based LLM agent pipeline that reduced token usage and inference cost by approximately 50% while preserving reasoning quality.
- Designed distillation methods to transfer reasoning and tool-use capabilities from Claude 3.7 to smaller language models, including Qwen3-4B, achieving performance comparable to the teacher model.

SELECTED PUBLICATIONS

- **Yuanpu Cao**, Saket Sathe, Hanyu Wang, Ziyi Yin, Fenglong Ma, Jinghui Chen
Progress-Aware Distillation for Mitigating Stagnation in Small Language Model Agents
Proceedings of the 40th Conference on Neural Information Processing Systems (NeurIPS) 2026
- **Yuanpu Cao**, Atharva Tendle, Rajarshi Bhowmik, Mounica Maddela, D. Preotiuc-Pietro, Lingjue Xie, Oshin Agarwal
Topology-Aware Process Supervision for Robust and Self-Improving Multi-Hop Reasoning
Under review
- **Yuanpu Cao**, Yongkang Du, Yurui Chang, Lu Lin, Jinghui Chen
From Narrative to Auditable Forecasts: A Structured Scaffold for Agentic Forecasting
Findings of the 2026 Conference on Empirical Methods in Natural Language Processing (EMNLP Findings) 2026

- **Yuanpu Cao**, Ziyi Yin, Fenglong Ma, Jinghui Chen
Can Factual Opinions Be Edited (Manipulated) in Large Language Models?
Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL) 2026
- **Yuanpu Cao**, Tianrong Zhang, Bochuan Cao, Ziyi Yin, Lu Lin, Fenglong Ma, Jinghui Chen
Personalized Steering of Large Language Models: Versatile Steering Vectors Through Bi-directional Preference Optimization
Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS) 2024
- **Yuanpu Cao**, Bochuan Cao, Jinghui Chen
Stealthy and Persistent Unalignment on Large Language Models via Backdoor Injections
Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) 2024
- Zehao Liu, **Yuanpu Cao**, Jinghui Chen, Vasant G Honavar
Restoring the Sweet Spot: Pass-Rate Weighted Self-Distillation for LLM Reasoning
Under review
- Yurui Chang, Yongkang Du, **Yuanpu Cao**, Jinghui Chen, Lu Lin
ForecastCompass: Guiding Agentic Forecasting with Adaptive Factor Memory
Proceedings of the 40th Conference on Neural Information Processing Systems (NeurIPS) 2026
- Hanyu Wang, Bochuan Cao, **Yuanpu Cao**, Jinghui Chen
TruthFlow: Truthful LLM Generation via Representation Flow Correction
Proceedings of the 42nd International Conference on Machine Learning (ICML) 2025
- Yaopei Zeng, **Yuanpu Cao**, Bochuan Cao, Yurui Chang, Jinghui Chen, Lu Lin
AdvI2I: Adversarial Image Attack on Image-to-Image Diffusion models
Proceedings of the 42nd International Conference on Machine Learning (ICML) 2025

TECHNICAL SKILLS

- **Programming Languages:** Python, C++, MATLAB
- **Machine Learning Frameworks:** PyTorch, TensorFlow 2
- **Web Framework:** Django
- **Databases:** MySQL, Redis