ZHAORUN CHEN

University of Chicago, 5801 S Ellis Ave, Chicago, IL 60637, USA
Homepage: billchan226.github.io | zhaorun@uchicago.edu

EDUCATION

Shanghai Jiao Tong University (SJTU) School of Electronic Information and Electrical Engineering（SEIEE） **Shanghai, China**
Bachelor of Engineering in Automation *Sept. 2018 – June 2022*
Purdue University Elmore Family School of Electrical and Computer Engineering **West Lafayette, IN, US**
Master in Electrical and Computer Engineering (full-funded) *Aug. 2022 – Aug. 2023*
University of Chicago Department of Computer Science **Chicago, IL, US**
PhD in Computer Science *Sept. 2024 – June 2030*

SELECTED PUBLICATIONS

[1] **Chen, Z.,** Xiang, Z., Xiao, C., Song, D., & Li, B. (2024). AgentPoison: Red-teaming LLM Agents via Poisoning Memory or Knowledge Bases, in Proceeding of the Thirty-Eighth Conference on Neural Information Processing Systems (NeurIPS 2024), Vancouver, Canada, Dec 2024.

[2] **Chen, Z.,** Zhao, Z., Luo, H., Yao, H., Li, B., & Zhou, J. (2024). HALC: Object Hallucination Reduction via Adaptive Focal-Contrast Decoding, in Proceeding of the Forty-first International Conference on Machine Learning (ICML 2024), Vienna, Austria, July 2024.

[3] Zhou, Y., Fan, Z., Cheng, D., Yang, S., **Chen, Z.**, Cui, C., Wang, X., Li, Y., Zhang, L., & Yao, H. (2024). Calibrated Self-Rewarding Vision Language Models, in Proceeding of the Thirty-Eighth Conference on Neural Information Processing Systems (NeurIPS 2024), Vancouver, Canada, Dec 2024.

[4] **Chen, Z.,** Zhao, Z., Zhu, Z., Zhang, R., Li, X., Raj, B., & Yao, H. (2024). AutoPRM: Automating Procedural Supervision for Multi-Step Reasoning via Controllable Question Decomposition, in Proceeding of 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2024), Mexico City, Mexico, Jun 2024.

[5] **Chen, Z.,** Wang, S., Zhao, Z., Mao, C., Zhou, Y., He, J., & Hu, A.S. (2024). ESCIRL: Evolving Self-Contrastive IRL for Trajectory Prediction in Autonomous Driving, in Proceeding of 8th Annual Conference on Robot Learning. (CoRL 2024), Munich, Germany, Nov 2024.

[6] **Chen, Z.**, Zhao, Z., He, T., Chen, B., Zhao, X., Gong, L., & Liu C. (2024). Safe Reinforcement Learning via Hierarchical Adaptive Chance-Constraint Safeguards, in Proceeding of 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2024), Abu Dhabi ,UAE, October 2024.

[7] **Chen, Z.,** Du, Y., Wen, Z., Zhou, Y., Cui, C., Weng, Z., ... , Rafailov, R., Finn, C., & Yao, H. (2024). MJ-Bench: Is Your Multimodal Reward Model Really a Good Judge for Text-to-Image Generation? In ICML 2024 Workshop on Foundation Models in the Wild, Vienna, Austria, July 2024.

RESEARCH EXPERIENCE

AgentPoison: Red-teaming LLM Agents via Poisoning Memory or Knowledge Bases Feb. 2024 – June 2024
Advisor: Prof. Bo Li, CS, UIUC/University of Chicago

- Propose the first backdoor attack framework targeting generic and RAG-based LLM agents by poisoning their long-term memory or knowledge base; propose an integral trigger optimization algorithm to map triggered instances to unique embedding space to achieve high backdoor attack success rate while preserving agent utility in non-triggered cases;
- On three types of real-world LLM agents (e.g. autonomous driving agent, QA agent, healthcare EHRAgent), AgentPoison achieves an average ASR $\geqslant$ 80% with minimal impact on benign performance ($\leqslant$ 1%) with a minor poison ratio < 0.1%.

HALC: Object Hallucination Reduction via Adaptive Focal-Contrast Decoding Aug. 2023 – Jan. 2024
Advisor: Prof. Bo Li, CS, University of Chicago

- Propose the key pattern to mitigating object hallucination is identifying the optimal visual context to safely predict each token;

propose an adaptive decoding algorithm HALC which employ a sampling based interpolated JSD-filter method to select proxies to approximate the optimal visual context and then perform contrastive decoding to reveal their context-specific information.

- HALC outperforms all SOTA (e.g. OPERA, VCD, woodpecker, LURE) in terms of CHAIR, POPE and MME benchmarks.

Automating Procedural Supervision for LLMs Reasoning via Controllable Problem Decomposition May 2023–Aug. 2023

Advisor: Prof. Huaxiu Yao, CS, UNC Chapel Hill; Prof. Bhiksha Raj, CS, Carnegie Mellon University

- Propose AutoPRM, a process-supervision pipeline that automatically decomposes long-form reasoning into sub-questions and assign fine-grained RL rewards to intermediate steps to enhance internal logic consistency and transfer of meta-knowledge.
- To allow fine-grained control during question solving, we propose a user-controlled switch to manage decomposition granularity and introduce FIM-guided-decoding to guide each subproblem solver towards solution of the primary problem. AutoPRM improves the accuracy on GSM8K (+4.4%), MATH (+4.2%) and StrategyQA (+6.2%) tasks over SOTA without external data.

INTERNSHIP EXPERIENCE

Safety Alignment and Guardrails for Video Generative Models | Research Scientist Intern May 2024 – Present

Advisor: Prof. Bo Li, VirtueAI

Self-rewarding Fine-grained Feedback for LLMs RLHF | Research Intern June 2023 – Feb. 2024

Advisor: Prof. Huaxiu Yao, UNC-Chapel Hill

Motion Planning for Decentralized Multi-manipulator Assembly System | AI Intern Feb. 2022 - Aug. 2022

Advisor: Prof. Cewu Lu, Flexiv Robotics Inc.

SKILLS

- **English Standard Test:** TOEFL: 111 (R: 28, L: 27, S: 28, W: 28)
- **Programming Languages:** solid expertise in Python, Matlab, C++, and various algorithms and data structures
- **Tools:** PyTorch, Matlab, LaTeX, TensorFlow, OpenCV, ROS, Linux, QT5, MySQL

AWARDS & HONOR

Full-funded two-year Guaranteed Fellowship at Purdue ECE (Very Rare) 2022

Best paper award in IEEE ISRIMT 2021 2021

Third Prize in the Contemporary Undergraduate Mathematical Contest in Modeling *2020*

First Prize in RoboCup China Open *2020*

Academic Excellence Scholarship (Top 5%) *2019-2022*

SERVICE

Graduate Teaching Assistant for Data Structure (ECE368) *2022-2023*

Conference Reviewer: NeurIPS'24, ICLR'24,25, AISTATS'24, COLM'24, ARR'24, IROS'24