ZHAORUN CHEN

(+1) 773-952-0790 Email: zhaorun@uchicago.edu ⋄ Homepage ⋄ Google Scholar

EDUCATION

University of Chicago 2024.09 - now
Ph.D. in Computer Science
– Advisor: Prof. Bo Li

Purdue University 2022.08 - 2023.08
M.S in Computer Engineering

Shanghai Jiao Tong University 2018.09 - 2022.06
B.E in Computer Engineering

PUBLICATIONS & PREPRINTS

Full publication list is in Google Scholar.

[1] **Zhaorun Chen**, Zhen Xiang, Chaowei Xiao, Dawn Song, Bo Li, 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. [Paper] [Code] [LLM Agent Safety]

[2] Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Sihan Yang, **Zhaorun Chen**, Chenhang Cui, Xiyao Wang, Yun Li, Linjun Zhang, Huaxiu Yao, Calibrated Self-Rewarding Vision Language Models, in Proceeding of *the Thirty-Eighth Conference on Neural Information Processing Systems (NeurIPS 2024)*, Vancouver, Canada, Dec 2024. [Paper] [Code] [RL Post-Training]

[3] **Zhaorun Chen**, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, Jiawei Zhou, 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. [Paper] [Code] [Hallucination]

[4] **Zhaorun Chen**, Zhuokai Zhao, Zhihong Zhu, Ruiqi Zhang, Xiang Li, Bhiksha Raj, Huaxiu Yao, 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. [Paper] [Code] [RL Post-Training]

[5] **Zhaorun Chen**, Francesco Pinto, Minzhou Pan, Bo Li, SafeWatch: An Efficient Safety-Policy Following Video Guardrail Model with Transparent Explanations, in *Under Review*. [Paper] [Video Generation Safety]

[6] Zhihong Zhu, Kefan Shen, **Zhaorun Chen**, Yunyan Zhang, Yuyan Chen, Xiaoqi Jiao, Zhongwei Wan, Shaorong Xie, Wei Liu, Xian Wu, Yefeng Zheng, DGLF: A Dual Graph-based Learning Framework for Multi-modal Sarcasm Detection, in Proceeding of *2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*, Miami, Florida, Nov 2024. [Multimodal Safety]

[7] **Zhaorun Chen**, Yichao Du, Zichen Wen, Yiyang Zhou, Chenhang Cui, Zhenzhen Weng, Haoqin Tu, Chaoqi Wang, Zhengwei Tong, Qinglan Huang, Canyu Chen, Qinghao Ye, Zhihong Zhu, Yuqing Zhang, Jiawei Zhou, Zhuokai Zhao, Rafael Rafailov, Chelsea Finn, Huaxiu Yao, 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*. [Paper] [Code] [RL Post-Training]

[8] Haibo Tong, Zhaoyang Wang, **Zhaorun Chen**, Haonian Ji, Shi Qiu, Siwei Han, Zhongkai Xue, Yiyang Zhou, Peng Xia, Kexin Geng, Mingyu Ding, Rafael Rafailov, Chelsea Finn, Huaxiu Yao, MJ-Bench-Video: A Fine-Grained Preference Dataset for Evaluating Reward Model of Text-to-Video Generation, in *Under Review*. [RL Post-Training]

[9] Siyue Wang, **Zhaorun Chen**, Zhuokai Zhao, Chaoli Mao, Yiyang Zhou, Jiayu He, Albert Sibo Hu, 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. [Paper] [Code] [RL for Robotics]

[10] **Zhaorun Chen**, Zhuokai Zhao, Tairan He, Binhao Chen, Xuhao Zhao, Liang Gong, Chengliang Liu, 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. [Paper] [Code] [RL for Robotics]

RESEARCH EXPERIENCES

Analyzing and Enhancing the Trustworthiness of LLM Agents via Red-teaming 2024
Advisor: Prof. Bo Li, University of Chicago/UIUC; Prof. Dawn Song, University of Berkeley

- Propose the **first backdoor attack** against generic LLM agents by poisoning their long-term memory or knowledge base; propose a trigger optimization algorithm to achieve high backdoor attack success rate ($\geq$ 95%) while preserving agent utility in non-triggered case ($\leq$ 1%), by injecting **only one** malicious memory;
- Our work AgentPoison [1] is published in NeurIPS 2024; Our approach can be practically used to attack ChatGPT and this vulnerability has recently been fixed by OpenAI in ChatGPT version 1.2024.247 [Link].

Certified Hallucination Reduction for LLMs via Sampling-time Intervention 2023 - 2024
Advisor: Prof. Bo Li, University of Chicago/UIUC

- First propose to address MLLMs hallucination through **grounded visual prompting** where we use a certified sampling-based algorithm to approximate the optimal visual context when decoding each token;
- Our work HALC [3] is published in ICML 2024, and has served as a standard MLLM decoding baseline in many subsequent hallucination-related research papers [Link].

Reward Modeling and RL Post-training for Text-to-Image generation 2024
Advisor: Prof. Chelsea Finn, Stanford University; Prof. Huaxiu Yao, UNC Chapel Hill

- Propose the first platform MJ-Bench [Link] to benchmark **multimodal reward models for text-to-image generation** and introduce a standard RLHF recipe for post-training multimodal foundation models.

Automating Post-Training Procedural Supervision for LLM Reasoning 2023 - 2024
Advisor: Prof. Huaxiu Yao, UNC Chapel Hill; Prof. Bhiksha Raj, CMU LTI

- Propose the first framework to improve the efficiency of training **procedural reward models** via problem decomposition to supervise LLM reasoning in long-form math problems (achieving SOTA in GSM8K/MATH);
- Our work AutoPRM [4] is published in NAACL 2024, and has inspired a lot of subsequent papers to efficiently automate training PRMs for fine-grained credit assignment in RLHF [Link].

INTERNSHIP EXPERIENCES

Virtue AI 2024
Mentor: Prof. Bo Li, Virtue AI CEO

- Develop the first video guardrail model SafeWatch [Link] and a 2M video safety benchmark SafeWatch-Bench to address the risks brought by the rapid proliferation of video generative models.
- SafeWatch is designed to handle a comprehensive collection of unsafe video scenarios and beat GPT-4o by 30% on average. Notably, SafeWatch can provide in-depth explanations for its guardrail decisions and incorporates multiple novel modules to handle the limitations in previous works, such as high inference latency and policy bias given long policy guidelines input.

ACADEMIC SERVICES

Conference Reviewer

- NeurIPS, ICLR, COLM, ARR, IROS 2024
- ICLR, CVPR 2025

Co-Organizer

- NeurIPS CLAS 2024: The Competition for LLM and Agent Safety [Link] 2024

Teaching Assistant

- Data Structure [Link] 2022-2023