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**, Mintong Kang, Bo Li, ShieldAgent: Shielding Agents via Verifiable Safety Policy Reasoning, in Proceedings of the Forty-second International Conference on Machine Learning **(ICML 2025)**, Vancouver, Canada, July 2025. [Paper] [Code] [Agent Safety] [Guardrail]

[2] **Zhaorun Chen**, Francesco Pinto, Minzhou Pan, Bo Li, SafeWatch: An Efficient Safety-Policy Following Video Guardrail Model with Transparent Explanations, in Proceedings of the 13th International Conference on Learning Representations **(ICLR 2025)**, Singapore, Apr 2025. [Paper] [Code] [Video Safety Reasoning] [RL Post-Training]

[3] **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]

[4] **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] [Multi-modal Hallucination]

[5] Chejian Xu, Jiawei Zhang, **Zhaorun Chen**, Chulin Xie, Mintong Kang, Zhuowen Yuan, Zidi Xiong, Chenhui Zhang, Lingzhi Yuan, Yi Zeng, Peiyang Xu, Chengquan Guo, Andy Zhou, Jeffrey Ziwei Tan, Zhun Wang, Alexander Xiong, Xuandong Zhao, Yu Gai, Francesco Pinto, Yujin Potter, Zhen Xiang, Zinan Lin, Dan Hendrycks, Dawn Song, Bo Li, MMDT: Decoding the Trustworthiness and Safety of Multimodal Foundation Models, in Proceedings of the 13th International Conference on Learning Representations **(ICLR 2025)**, Singapore, Apr 2025. [Paper] [Code] [Multi-modal Safety]

[6] Peng Xia, Siwei Han, Shi Qiu, Yiyang Zhou, Zhaoyang Wang, Wenhao Zheng, **Zhaorun Chen**, Chenhang Cui, Mingyu Ding, Linjie Li, Lijuan Wang, Huaxiu Yao, MMIE: Massive Multimodal Interleaved Comprehension Benchmark for Large Vision-Language Models, in Proceedings of the 13th International Conference on Learning Representations **(ICLR 2025)** (**Oral Presentation**), Singapore, Apr 2025. [Paper] [Code] [Multi-modal Reasoning]

[7] Yiyang Zhou, Zhaoyang Wang, Tianle Wang, Shangyu Xing, Peng Xia, Bo Li, Kaiyuan Zheng, Zijian Zhang, **Zhaorun Chen**, Wenhao Zheng, Xuchao Zhang, Chetan Bansal, Weitong Zhang, Ying Wei, Mohit Bansal, Huaxiu Yao, AnyPrefer: An Automatic Framework for Preference Data Synthesis, in Proceedings of the 13th International Conference on Learning Representations **(ICLR 2025)**, Singapore, Apr 2025. [Paper] [RL Post-Training]

[8] Chenhang Cui, An Zhang, Yiyang Zhou, **Zhaorun Chen**, Gelei Deng, Huaxiu Yao, Tat-Seng Chua, Fine-Grained Verifiers: Preference Modeling as Next-token in Vision-Language Alignment, in Proceedings of the 13th International Conference on Learning Representations **(ICLR 2025)**, Singapore, Apr 2025. [Paper] [RL Post-Training]

[9] 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]

[10] **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]

[11] 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]

[12] **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 *Under Review*. [Paper] [Code] [RL Post-Training]

[13] 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]

[14] 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]

[15] **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]

WORK EXPERIENCES

Meta Superintelligence Labs Summer, 2025
Mentor: Dat Huynh, Meta Superintelligence Labs

- To bridge the high cost of training agents via online RL in expensive real-world environments, we propose a synthetic agent RL framework by training a world model that co-evolves with the LLM agent to continuously synthesize challenging trajectories for training the agent.

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

- Develop the first video guardrail model SafeWatch [Link] (accepted by ICLR 2025) and a 2M video safety benchmark SafeWatch-Bench to address the risks brought by recent powerful video generative models.

RESEARCH EXPERIENCES

Safeguarding General AI Agents via Safety Policy Reasoning 2025
Advisor: Prof. Bo Li, University of Chicago/UIUC/Virtue AI

- Propose the **first agentic guardrail system** ShieldAgent designed to enforce explicit safety policy compliance for other AI agents' action trajectories through logical reasoning; introduce the first agent guardrail dataset ShieldAgent-Bench collected via SOTA attacks across 6 web environments and 7 risk categories;
- ShieldAgent [1] is published in **ICML 2025**; it achieves SOTA on ShieldAgent-Bench and three existing agent safety benchmarks, outperforming prior methods by 11.3% on average with a high recall of 90.1%.

An Efficient Video Guardrail Model for Safeguarding Video Generative Models 2025
Advisor: Prof. Bo Li, University of Chicago/UIUC/Virtue AI

- Propose the **first video guardrail model** SafeWatch which can efficiently follow customized safety policies and provide multi-label video guardrails with content-specific explanations; introduce a 2M video safety dataset SafeWatch-Bench that covers over 30 unsafe video scenarios for training and benchmark SafeWatch;
- SafeWatch [2] is published in **ICLR 2025**; it outperforms SOTAs on both real-world and generative subset of SafeWatch-Bench by 28.2%; also achieves SOTA on 5 existing benchmarks and 8 unseen video categories.

Analyzing the Safety of RAG-based LLM Agents via Poisoning Attack 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 [3] 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 [4] 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 [10] 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].

ACADEMIC SERVICES

Conference Area Chair

– EMNLP 2025

Conference Reviewer

– NeurIPS, ICLR, COLM, ARR, IROS 2024
– ICLR, CVPR, ICML, NeurIPS 2025

Organizer

– NeurIPS CLAS 2024: The Competition for LLM and Agent Safety [Link] 2024
– COLM 2025 Workshop on AI Agents: Capabilities and Safety (AIA 2025) [Link] 2025