Hanjun Luo

Abu Dhabi, UAE — hl6266@nyu.edu — Google Scholar — GitHub — Homepage

EDUCATION

Ph.D. in Computer Science
New York University
Expected completion: 2030

Bachelor of Engineering in Computer Engineering
Zhejiang University
Year of completion: 2025

Bachelor of Science in Computer Engineering
University of Illinois Urbana-Champaign
Year of completion: 2025

PUBLICATIONS

Li, K., Shen, C., Liu, Y., Han, J., Zheng, K., Zou, X., Wang, Z., Du, X., Zhang, S., **Luo, H.**, *et al.* (2026). AudioTrust: Benchmarking the multifaceted trustworthiness of audio large language models. *International Conference on Learning Representations (ICLR 2026), poster.*

Luo, H., Dai, S., Ni, C., Li, X., Zhang, G., Wang, K., Liu, T., & Salam, H. (2025). AgentAuditor: Human-level safety and security evaluation for LLM agents. *Conference on Neural Information Processing Systems (NeurIPS 2025), poster.*

Luo, H., Jin, Y., Li, X., Liu, X., Chen, R., Shang, T., Wang, K., Wen, Q., & Liu, Z. (2025). DynamicNER: A dynamic, multilingual, and fine-grained dataset for LLM-based named entity recognition. *Conference on Empirical Methods in Natural Language Processing (EMNLP 2025), main conference.*

Luo, H., Deng, Z., Chen, R., & Liu, Z. (2024). FAIntbench: A holistic and precise benchmark for bias evaluation in text-to-image models. *ICML 2024 Workshop on Data-centric Machine Learning Research (DMLR).*

Sun, M., Zhao, Z., Chai, W., **Luo, H.**, Cao, S., Zhang, Y., Hwang, J.-N., & Wang, G. (2024). UniAP: Towards universal animal perception in vision via few-shot learning. *AAAI Conference on Artificial Intelligence (AAAI 2024).*

PREPRINTS

Li, J., Chen, Z., **Luo, H.**, & Salam, H. (2026). PrefIx: Understand and Adapt to User Preference in Human-Agent Interaction. *arXiv:2602.06714.*

Luo, H., Ni, C., Wen, J., Huang, Z., Wang, Y., Liao, B., Chung, S., Jin, Y., Li, X., Xu, W., Wang, X., & Salam, H. (2025). HAI-Eval: Measuring human-AI synergy in collaborative coding. *arXiv:2512.04111.*

Cai, H., Rahman, M. M., Dong, M., Li, J., Pu, M., Fang, Z., Peng, Y., **Luo, H.**, & Liu, Y. (2025). AutoDebias: Automated framework for debiasing text-to-image models. *arXiv:2508.00445.*

Wang, K., Zhang, G., Zhou, Z., Wu, J., Yu, M., Zhao, S., Yin, C., Fu, J., Yan, Y., **Luo, H.**, *et al.* (2025). A comprehensive survey in LLM(-agent) full stack safety: Data, training and deployment. *arXiv:2504.15585.*

Luo, H., Deng, Z., Huang, H., Liu, X., Chen, R., & Liu, Z. (2024). VersusDebias: Universal zero-shot debiasing for text-to-image models via SLM-based prompt engineering and generative adversary. *arXiv:2407.19524.*

Luo, H., Huang, H., Deng, Z., Liu, X., Chen, R., & Liu, Z. (2024). BIGbench: A unified benchmark for social bias in text-to-image generative models based on multi-modal LLM. *arXiv:2407.15240.*

RESEARCH EXPERIENCE

LLM Agent Safety
2024.11–present
New York University Abu Dhabi & Nanyang Technology University, Prof. Hanan Salam, Prof. XiaoFeng Wang, Prof. Yang Liu, Dr. Xinfeng Li
– Representative works: AgentAuditor (NeurIPS 2025) and a full-stack safety survey for LLM(-agent) systems.

Human–Agent Collaboration
2025.11–present
New York University Abu Dhabi, Prof. Hanan Salam
– Representative works: HAI-Eval and PrefIx.

Bias and Debiasing in Text-to-Image Models
2024.4–present
Zhejiang University & Nanyang Technology University, Prof. Zuozhu Liu, Dr. Xinfeng Li
– Representative works: FAIntbench (ICML-DMLR 2024), BIGbench, VersusDebias, and AutoDebias.

Audio Model Trustworthiness
2024.11–present
Nanyang Technology University, Prof. XiaoFeng Wang & Dr. Xinfeng Li
– Representative works: AudioTrust (ICLR 2026), AudioStealer.

LLM-based Multilingual Information Extraction
2024.7–present
Zhejiang University, Prof. Zuozhu Liu
– Representative work: DynamicNER (EMNLP 2025).

Computer Vision
2023.7–2023.11
Zhejiang University, Prof. Gaoang Wang
– Representative work: UniAP (AAAI 2024).

TEACHING EXPERIENCE

CS101: Intro to Computing
2024.9–2024.12
Teaching Assistant, Zhejiang University, Prof. Wee Liat Ong

ADDITIONAL SKILLS

Languages
– Chinese (native), English (fluent)

Software Development
– Completed core courses in machine learning, computer systems, data structures, algorithms

Photography
– Signed contributor for Visual China Group (VCG); experienced in documentary and landscape photography

RESEARCH INTERESTS

- Safety and security of LLM agents
- Human-agent collaboration
- Bias and trustworthiness in generative and multimodal models
- LLM-based multilingual information extraction