Introduction



Haowen Pan is currently pursuing the Ph.D. degree with the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, under the supervision of Prof. Xun Yang. He received the B.S. degree in Artificial Intelligence and Data Science from USTC in 2023. His research interests include natural language processing, multimodal information processing, model interpretability, and model security.

Google Scholar Homepage
Email: phw1129@mail.ustc.edu.cn Phone: +86 18575606474

Research Summary

Haowen Pan is enthusiastic about understanding and improving the trustworthiness of large language models (LLMs). His research aims to study how LLMs acquire, represent, and utilize knowledge, as well as how these processes can be made more reliable and secure. His current research interests include LLM knowledge editing, multimodal model interpretability, memory security of LLM agents, and LLM system security. Ultimately, he seeks to develop AI systems that are more reliable, controllable, safe, and responsible for real-world deployment.

Selected Publications

*** denotes the corresponding author, † denotes equal contribution.**

Under Review:

- **Haowen Pan**, Pengyang Shao, Yunshan Ma, Yanzheng Jin, Naixin Zhai, Xiang Wang, Xun Yang. MemForm: Benchmarking Memory Formation Poisoning in LLM Agents. Submitted to The Fifteenth International Conference on Learning Representations (ICLR 2027).
- **Haowen Pan**, Xun Yang, Rong Dai, Yonggang Zhang, Ming Pei, Tongliang Liu, Bo Han. Securing System Prompts from Unauthorized Leakage in Third-Party LLM Services. Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI).

Published:

- Xun Yang, **Haowen Pan**, Xiaozhi Wang, Yixin Cao*, Juanzi Li, Meng Wang. FiNE: Fine-grained Neuron-level Model Editing for Reliable and Safe LLMs. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2026. [PDF] [Code]
- **Haowen Pan**, Xiaozhi Wang, Yixin Cao*, Zenglin Shi, Xun Yang*, Juanzi Li, Meng Wang. Precise Localization of Memories: A Fine-grained Neuron-level Knowledge

Editing Technique for LLMs. The Thirteenth International Conference on Learning Representations (ICLR 2025). [PDF] [Code]

- **Haowen Pan**, Yixin Cao, Xiaozhi Wang, Xun Yang*, Meng Wang. Finding and Editing Multi-Modal Neurons in Pre-Trained Transformers. Findings of the Association for Computational Linguistics (ACL 2024). [PDF] [Code]

Experiences

1)	University of Science and Technology of China Advisor: M.S. (2023–2025), transferred to Ph.D.	Ph.D. Xun Yang	2025 – Present Hefei, China
2)	University of Science and Technology of China	B.S.	2019 – 2023 Hefei, China
3)	iFLYTEK Co., Ltd. Advisor:	Intern Ya Li	2022 Hefei, China