

董建硕 (Jianshuo Dong)

📞 18522413148 ✉️ dongjs23@mails.tsinghua.edu.cn



Education

Tsinghua University	2023.09 – Now
PhD Candidate Institute of Network Science and Cyberspace Advisor: Han Qiu	Beijing
Wuhan University	2019.09 – 2023.06
Bachelor School of Cyber Science and Engineering Supervisors: Qian Wang, Yanjiao Chen	Wuhan

Research Interests

- **Safety & Security of Large Model Systems:** Deliver trustworthy large model services to users.
- **Autonomous Agentic AI:** Developing strong, safe, and efficient autonomous agents.
- **Explainable AI:** Reverse-engineer and closely monitor AI models.

Published Papers

Revisiting the Reliability of Language Models in Instruction-Following

ACL 2026 **Oral** CCF-A/TH-CPL-A

Jianshuo Dong, Yutong Zhang, Yan Liu, Zhenyu Zhong, Tao Wei, Chao Zhang, Han Qiu

Evaluating the reliability of LLMs in instruction-following

SafeSearch: Automated Red-Teaming of LLM-Based Search Agents

ICML 2026 CCF-A/TH-CPL-A

Jianshuo Dong, Sheng Guo, Hao Wang, Xun Chen, Zhuotao Liu, Tianwei Zhang, Ke Xu, Minlie Huang, Han Qiu

Automated red-teaming framework for LLM-based search agents

“I’ve Decided to Leak”: Probing Internals Behind Prompt Leakage Intents

EMNLP 2025 **Oral** TH-CPL-A

Jianshuo Dong, Yutong Zhang, Yan Liu, Zhenyu Zhong, Tao Wei, Ke Xu, Minlie Huang, Chao Zhang, Han Qiu

Explaining prompt leakage via probing

An Engorgio Prompt Makes Large Language Model Babble on

ICLR 2025 TH-CPL-A

Jianshuo Dong, Ziyuan Zhang, Qingjie Zhang, Tianwei Zhang, Hao Wang, Hewu Li, Qi Li, Chao Zhang, Ke Xu, Han Qiu

Attacking availability of LLM systems

One-bit Flip Is All You Need: When Bit-Flip Attack Meets Model Training

ICCV 2023 CCF-A/TH-CPL-A

Jianshuo Dong, Han Qiu, Yiming Li, Tianwei Zhang, Yuanjie Li, Zeqi Lai, Chao Zhang, Shu-Tao Xia

Targeted bit flip attack

Academic Services

- Teaching Assistant: Trustworthy Machine Learning, Tsinghua University, Spring 2025.
- ICLR Reviewer: 2025 (**Notable**), 2026.
- NeurIPS Reviewer: 2025 (**Top**).
- ICML Reviewer: 2026 (**Gold**).