

EDUCATION

University of Science and Technology of China	Sep. 2025 – Jun. 2028 (expected)
M.E. in Artificial Intelligence	Hefei, China
- Supervisors: Prof. Haisheng Tan and Prof. Haohua Du - Research Focus: LLM/Agent Safety - Coursework: Mathematical Statistics, Machine Learning, Reinforcement Learning, ML Systems	
Beihang University	Sep. 2021 – Jun. 2025
B.E. in Cyberspace Security	Beijing, China
- Supervisors: Prof. Haohua Du - GPA: 3.81 / 4.0 - Coursework: Probability and Statistics, Computation Theory, Advanced Algebra, Information Theory	

RESEARCH INTERESTS

I am deeply motivated to contribute to LLM and Agent safety research, particularly in developing robust methods to detect and mitigate jailbreaks, prompt injection, and reasoning failures.

PUBLICATIONS

Authors in bold denote me. * equal contribution. † corresponding author.

- Z. Liu**, H. Du[†], H. Sun and H. Tan[†]. *Fine-Grained Selective Refusal: Steering LLM Safety via Wasserstein-Informed Flow Matching*. EMNLP 2026 Main Conference.
- Z. Liu**, H. Sun, H. Du[†], Z. Wang and H. Tan. *HALO: Hyperspherical Information Geometry for LLM Self-Evaluation*. Submitted to AAAI 2027.
- W. Ma, **Z. Liu**, H. Du[†] and X. Liang. *Sedative: Pacify the Online Social Networks under Hijacking-based Troll Attacks*. IWQoS 2025, CCF-B.
- Y. Yao[†], Y. Zhang, **Z. Liu**, Y. Wang, X. Tan and Z. Zhai. *Traceable and revocable multi-authority ABE supporting decryption outsourcing and policy update for cloud data access control*. *Frontiers of Computer Science*, CCF-B.

Patents

- A public opinion control method against hijacking-style malicious provocative attacks.
- Efficient Multi-functional Access Control via Attribute-Based Encryption.

INTERNSHIPS

Research Intern, BraneMatrix	Jul. 2026 – Jan. 2027 (expected)
Mentor: Dr. Xiaojun Jia (NTU)	Shanghai, China

AWARDS & HONORS

First Prize of Academic Scholarship, USTC	2025
Outstanding Graduate, Beihang University	2025
National Scholarship, Ministry of Education of China	2024
Merit Student, Beihang University	2023, 2024

SKILLS

Languages	English (CET-6), Mandarin (native)
Programming	Python, C, C++, MATLAB
Frameworks & Tools	PyTorch, Git, Docker