Hao li （李昊）

hao612@buaa.edu.cn · Beihang LLLeoLi (+86) 13844640268 ·

Education

Beihang University, Beijing, China 2024 – Present

Master student in School of AI, advised by *Prof. Lei Sha*, expected Jun. 2027

Beihang University, Beijing, China 2020 – 2024

B.Eng. in School of Software, GPA 3.85/4.00

Experience

1. **Research Intern @ JD.com** Jul. 2026 – Present
   Agentic RL Mentor: Keqing He
2. **Research Intern @ TikTok** Mar. 2026 – Jun. 2026
   Agentic RL Mentor: Longtao Zheng
3. **Research Assistant @ HKUST** Jan. 2026 – Present
   Agentic RL Advisor: Prof. Junxian He
4. **Research Intern @ BAAI** Aug.2025 – Dec. 2025
   MLLM Evaluation Mentor: Jinge Yao
5. **Research Intern @ Shanghai AI Lab** Jul. 2024 – Jul. 2025
   AI Safety Advisor: Jing Shao
6. **Research Intern @ CoAI (THU)** Oct. 2023 – Jan. 2025
   LLM Safety Advisor: Minlie Huang

Papers

1. **ProgAgent: Learning Tool Orchestration through Programmatic Training** [Paper]
   ICLR 2027 Under Review First Author
   TL;DR: ProgAgent trains agents to orchestrate tools through programmatic tool calling.
2. **SafeSteer: Localized On-Policy Distillation for Efficient Safety Alignment** [Paper]
   EMNLP 2026 Oral (3%) First Author
   TL;DR: SafeSteer distills safety behaviors into selected tokens while preserving general capabilities.
3. **Layer-Aware Representation Filtering: Purifying Finetuning Data to Preserve LLM Safety Alignment** [Paper]
   EMNLP 2025 Main First Author
   TL;DR: LARF filters safety-degrading data using representations from safety-sensitive layers.
4. **LLMs know their vulnerabilities: Uncover Safety Gaps through Natural Distribution Shifts** [Paper]
   ACL 2025 Outstanding Paper (0.3%) Co-first Author, Student First Author
   TL;DR: ActorBreaker exposes LLM safety gaps through distribution shifts in multi-turn interactions.
5. **ASETF: A Novel Method for Jailbreak Attack on LLMs through Translate Suffix Embeddings** [Paper]
   EMNLP 2024 Oral (3%) Co-first Author
   TL;DR: ASETF translates adversarial suffix embeddings into fluent text for efficient LLM jailbreaking.

6. **Mine Odyssey: Benchmarking Spatial Agentic Intelligence in the Wild** [Paper]
   ICLR 2027 Under Review Second Author
   TL;DR: Mine Odyssey benchmarks spatial agentic intelligence in MC real-world location reconstructions.
7. **Be Your Own Red Teamer: Safety Alignment via Self-Play and Reflective Experience Replay** [Paper]
   ACL 2026 Findings Third Author
   TL;DR: SSP strengthens LLM defenses through adversarial self-play and reflective experience replay.
8. **Beyond Multiple Choice: Verifiable OpenQA for Robust Vision-Language RFT** [Paper]
   CVPR 2026 Second Author
   TL;DR: ReVeL converts multiple-choice QAs into verifiable open-ended QAs for robust VLM RL.
9. **DiffusionAttacker: Diffusion-Driven Prompt Manipulation for LLM Jailbreak** [Paper]
   EMNLP 2025 Oral (3%) Second Author
   TL;DR: DA uses diffusion to generate fluent, diverse jailbreaks through semantic-preserving rewriting.
10. **VLSBench: Unveiling Visual Leakage in Multimodal Safety** [Paper]
    ACL 2025 Main Third Author
    TL;DR: VLSBench evaluates VLM safety without leaking visual risk information through textual queries.

Projects

1. **Toolathlon** [Project]
   Github star **500+**| [ICLR 2026] The Tool Decathlon: Benchmarking Language Agents for Diverse, Realistic, and Long-Horizon Task Execution.
2. **thu-coai/AISafetyLab** [Project]
   Github star **250+**| A comprehensive framework covering safety attack, defense, evaluation and paper list.
3. **RAPTOR** [Project]
   Github star **1.8k+**| [ICLR2023] The official implementation of RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval.

Honors and Awards

National Scholarship 2025
Challenge Cup (National Special Award) 2023
Academic Excellence Scholarship (Beihang University) 2021-2025

Services

- Reviewer: COLM, ICLR, ACL, EMNLP, NAACL, AAAI, WWW
- In-School: TA in School of Software and School of AI, Beihang University