LAIZHEN LI

lilaizhencs@gmail.com github.com/Lilaizhen Google Scholar

EDUCATION

University of Chinese Academy of Sciences, China *Sep. 2024–Present (Expected Jul. 2027)*
Master of Engineering in Computer Technology GPA: 3.71/4.0
Advisor: Xitong Gao

Henan University of Engineering, China *Sep. 2019–Jul. 2023*
Bachelor of Engineering in Software Engineering Average: 81.73/100

RESEARCH INTERESTS

I work on large language models (LLMs) and AI agents, with a focus on safety, security, and efficiency.

RESEARCH EXPERIENCE

Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences *Jun. 2025–Present*

Master's Student *Advisor: Xitong Gao*

LLM Safety — StreamGuard

- Designed and implemented **StreamGuard**, a lightweight, plug-and-play decoding-time defense against jailbreak attacks that self-reviews partial generations and inserts safety prompts when unsafe content is detected, without fine-tuning.
- Evaluated StreamGuard on Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct against AutoDAN, GCG, and DeepInception; achieved a mean ASR of **1.7%**, compared with **20.3%** for the best-performing defense baseline and **60.3%** without defense, across six model–attack settings.

Agent Security — A2M

- Designed and implemented **A2M (Attraction to Manipulation)**, a two-stage black-box attack that optimizes tool metadata (tool names and descriptions) to increase selection likelihood, then refines return payloads using execution traces to manipulate downstream reasoning.
- Evaluated on LiveMCPBench across five frontier models; when optimized directly against GLM-4.6, A2M achieved a mean ASR of **74.4%** (64.3–92.9%) across three ASR-based objectives (information exfiltration, unauthorized environment modification, and reasoning derailment) and increased token consumption by **32.4**× on a fourth, token-wasting objective.

Agent Efficiency — Growing Harness

- Designed and implemented **Growing Harness**, a failure-guided framework that starts from a strategy-free scaffold and incrementally grows it into an executable agent harness through function-level execution traces and localized code edits, shifting recurring control logic from LLM inference to reusable program code.
- Evaluated on BrowseComp-Plus and WebArena-Verified across three backbone models; achieved a mean task success rate of **42.2%** across six benchmark–model settings, versus **23.6%** for the ReAct-style Tool-Calling baseline (**+18.6 pp**), while reducing LLM calls by **76.0–91.8%** and inference cost by **74.4–98.6%**.

PUBLICATIONS

[1] A2M: Trace-Optimized Agent Hijacking in the MCP Ecosystem
Laizhen Li, Xuan Wang, Peicheng Zhao, Juanjuan Zhao, Kejiang Ye, Cheng-zhong Xu, and Xitong Gao.
AACL-IJCNLP 2026 (Main Conference, Oral Presentation). **[Paper]**

[2] StreamGuard: A Streaming-based Defense Against Jailbreaking Attacks in Large Language Models
Laizhen Li, Xitong Gao, Xuan Wang, Juanjuan Zhao, and Kejiang Ye.
ICONIP 2025. **[Paper]**

[3] JADE: Jailbreak-Aware Dynamic Defense Enhancer for LLMs via Reinforcement Learning
Peicheng Zhao, **Laizhen Li**, Xitong Gao, Juanjuan Zhao, and Kejiang Ye.
IEEE BigData 2025. **[Paper]**

MANUSCRIPTS UNDER REVIEW

[1] Grow the Harness, Not the Context: From Strategy-Free Scaffolds to Reusable Specialist Agents
Laizhen Li, Jiarui Li, Juanjuan Zhao, Kejiang Ye, Ye Li, Cheng-zhong Xu, and Xitong Gao. **[Paper]**

SKILLS

Programming & Tools: Python, Bash, Linux, Git, Docker.

Machine Learning: PyTorch, Hugging Face Transformers, Weights & Biases.

LLM & Agents: OpenRLHF, vLLM, LangChain, MCP, SFT, PPO, LoRA.