Jiacheng Liang

814-470-0569 | ljcpro@outlook.com | LinkedIn | Google Scholar | jackpurcell.github.io

Research Interest

My research focuses on **LLM safety and trustworthy AI**, with an emphasis on identifying security vulnerabilities and developing robust defenses. Research areas include LLM post-training and safety alignment, LLM agent safety, LLM deployment security, jailbreak and backdoor defense, and RAG safety.

Education

Stony Brook University & Penn State University — Stony Brook, NY
Ph.D. Candidate in Computer Science, GPA: 4.0/4.0 — *Sep 2022 – Dec 2026 (Expected)*

University of Electronic Science and Technology of China — Chengdu, China
B.Eng. in Software Engineering - International Elite Class — *Sep 2018 – Jun 2022*

Publications

Peer-Reviewed Publications

[1] **Liang, J.**, Jiang, T., Wang, Y., Zhu, R., Ma, F., & Wang, T. (2025). AutoRAN: Automated Hijacking of Safety Reasoning in Large Reasoning Models. ***ACL 2026***. [pdf]

[2] **Liang, J.**, Ma, Y., Kumarage, T., Krishna, S., Gupta, R., Chang, K.-W., Galstyan, A., & Peris, C. (2026). ARES: Adaptive Red-Teaming and End-to-End Repair of Policy-Reward System. ***ACL 2026***. [pdf]

[3] **Liang, J.**, Wang, Y., Li, C., Zhu, R., Jiang, T., Gong, N., & Wang, T. (2025). GraphRAG under Fire. ***IEEE S&P 2026 (Top-1 Security Conference)***. [pdf]

[4] Jiang, T., **Liang, J.**, Zhu, R., Zhou, J., Ma, F., & Wang, T. (2026). Dynamic Token Reweighting for Robust Vision-Language Models. **CVPR 2026**. [pdf]

[5] Wang, Y., Chen, G., **Liang, J.**, Wang, T., & Jiang, T. (2026). Reasoning or Retrieval? A Study of Answer Attribution on Large Reasoning Models. **ICLR 2026**. [pdf]

[6] **Liang, J.**, Wang, Z., Hong, L., Ji, S., & Wang, T. (2025). WaterPark: A Robustness Assessment of Language Model Watermarking. ***EMNLP 2025***. [pdf]

[7] Liu, X.*, **Liang, J.***, Tang, L., You, C., Ye, M., & Xi, Z. (2025). Data to Defense: The Role of Curation in Customizing LLMs Against Jailbreaking Attacks. ***EMNLP 2025***. [pdf]

[8] Jiang, T., Wang, Z., **Liang, J.**, Li, C., Wang, Y., & Wang, T. (2025). RobustKV: Defending Large Language Models against Jailbreak Attacks via KV Eviction. ***ICLR 2025***. [pdf]

[9] **Liang, J.**, Pang, R., Li, C., & Wang, T. (2024). Model Extraction Attacks Revisited. ***AsiaCCS 2024***.

[10] **Liang, J.**, Li, S., Cao, B., Jiang, W., & He, C. (2021). Omnilytics: A Blockchain-based Secure Data Market for Decentralized ML. ***ICML Workshop on Federated Learning 2021***.

[11] Zhou, Q., Guo, S., Pan, J., **Liang, J.**, Guo, J., Xu, Z., & Zhou, J. (2024). PASS: Patch Automatic Skip Scheme for Efficient On-Device Video Perception. ***IEEE TPAMI***.

[12] Zhou, Q., Guo, S., Pan, J., **Liang, J.**, Xu, Z., & Zhou, J. (2023). PASS: Patch Automatic Skip Scheme for Efficient Real-Time Video Perception on Edge Devices. ***AAAI 2023***.

[13] Li, J., Wei, G., **Liang, J.**, Ren, Y., Lee, P. P., & Zhang, X. (2022). Revisiting Frequency Analysis Against Encrypted Deduplication via Statistical Distribution. ***IEEE INFOCOM 2022***.

Under Review / Preprints

[14] **Liang, J.**, Wang, Y., Jiang, T., & Wang, T. RASA: Routing-Aware Safety Alignment for Mixture-of-Experts Models. **Submitted to COLM 2026**. [pdf]

[15] Jiang, T., Wang, Y., **Liang, J.**, & Wang, T. (2026). AgentLAB: Benchmarking LLM Agents Against Long-Horizon Attacks. ***Submitted to ICML 2026***. [pdf]

[16] Liu, X.*, **Liang, J.***, Yan, Q., Jang, J., Mao, S., Ye, M., Jia, J., & Xi, Z. (2026). CyLens: Towards reinventing cyber threat intelligence in the paradigm of agentic large language models. [pdf]

[17] Zhu, R., Wang, Y., Jiang, T., **Liang, J.**, & Wang, T. (2025). Self-Improving Model Steering. [pdf]

[18] Li, C., **Liang, J.**, Cao, B., Chen, J., & Wang, T. (2025). Your Agent Can Defend Itself Against Backdoor Attacks. [pdf]

[19] Xu, N., Li, C., Du, T., Li, M., Luo, W., **Liang, J.**, Li, Y., Zhang, X., Han, M., Yin, J., & Wang, T. (2024). CopyrightMeter: Revisiting Copyright Protection in Text-to-Image Models. [pdf]

Experience

Amazon AGI Foundation (Responsible AI Team) Boston, MA

Applied Scientist Intern *May 2025 – Aug 2025*

- **Adaptive Red-Teaming for RLHF** : Identify systemic weaknesses in RLHF pipelines, where both the Core LLM and Reward Model fail in tandem, and propose an adaptive red-teaming + end-to-end repair framework for more robust alignment. (Pub. [2])

Recent Academic Projects

Security Challenges in Large Language Models Oct 2023 – Present

Research Assistant @ALPS-Lab (Prof. Ting Wang), Stony Brook University *Stony Brook, NY*

- **Data Poisoning in GraphRAG**: Identify and analyze potential security threats within GraphRAG, focusing on vulnerabilities that could compromise the integrity of the entity relationship graphs and the overall knowledge base. (Pub. [3])
- **Hijacking Reasoning Models**: Automated hijacking of internal safety reasoning in large reasoning models, showing that reasoning transparency itself can become an exploitable attack surface. (Pub. [1])
- **Long-Horizon Attack Benchmarking for LLM Agents**: Built AgentLAB, the first extensible benchmark for long-horizon security attacks on LLM agents, uncovering fundamental vulnerabilities beyond single-turn defenses. (Pub. [15])
- **LLM Watermark's Robustness Benchmarking**: Investigate existing LLM watermark methods to form a Systematization of Knowledge(SoK) and establish a comprehensive evaluation platform to standardize their robustness. Propose an advanced method to attack the watermark and discuss the corresponding defenses. (Pub. [6])

Defensive Strategies for Large Language Models Aug 2024 – Present

Research Assistant @ALPS-Lab (Prof. Ting Wang), Stony Brook University *Stony Brook, NY*

- **Mixture of experts (MoE) LLM Safety Alignment** : Developed a dual-layered safety alignment framework for Mixture-of-Experts (MoE) models using a bi-level alternating optimization strategy to robustify experts and routers against adversarial jailbreaks. (Pub. [14])
- **Jailbreak Defense Through Data Curation**: Created a defensive framework to mitigate jailbreaking attacks at every stage of LLM customization, achieving a 100% success rate in generating responsible responses. (Pub. [7])
- **Jailbreak Attack Prevention Through KV Eviction**: Designed a defense against jailbreak attacks by selectively evicting low-importance tokens from key-value caches, countering adversarial prompts while preserving LLM performance on benign queries. (Pub. [8])
- **Multimodal jailbreak defense**: Designed a defense mechanism against multimodal jailbreak attacks by reweighting visual tokens during inference. DTR is the first to apply KV cache optimization for robust multimodal LLM safety. (Pub. [4])
- **Backdoor Attack Defense for LLM Agents**: Developed a defense system that enables LLM agents to defend themselves against backdoor attacks by ensuring consistency between agent planning, execution, and user instructions. (Pub. [18])

Model Extraction Security Challenges on Real MLaaS APIs Aug 2022 – Oct 2023

Research Assistant @ALPS-Lab (Prof. Ting Wang), Stony Brook University *Stony Brook, NY*

- **Model Extraction Platform Development**: Designed and launched "MEBench", an easy-to-use open-source evaluation tool, assessing ME vulnerabilities in various MLaaS APIs by integrating multiple attacks, metrics, and models. (Pub. [9])