Jiacheng Liang

814-470-0569 | ljcpro@outlook.com | jiachliang@cs.stonybrook.edu | LinkedIn | Google Scholar | Github Page |Stony Brook, NY

Research Interest

My expertise is in ensuring the **Safety and Trustworthiness of Large Language Models (LLMs)**. I work on identifying **Security Challenges** and developing **Defensive Strategies** to protect these models from adversarial threats. My recent work includes:

- **Security Challenges:** Investigate vulnerabilities in **LLM watermarking and GraphRAG**; use long CoT to **jailbreak reasoning model** and propose advanced methods to address these weaknesses.
- **Defensive Strategies:** Develop methods to **defend against backdoor attacks (self-defense mechanisms in agent) and jailbreaking (Data curation when finetuning and KV eviction when inference).**

In addition, I possess extensive expertise and a strong interest in **post-training, prompt engineering, inference optimization, LLM agents, and ensuring LLM alignment.**

Education

Stony Brook University & Penn State University — Stony Brook, NY
Ph.D. candidate; Department of Computer Science; GPA: 4.0/4.0 — 09/2022 – 05/2026

University of Electronic Science and Technology of China — Chengdu, China
B.S. of Software Engineering - International Elite Class — 09/2018 – 06/2022

Skills

Knowledge: Large Language Model, LLM security, GraphRAG, RAG, LoRA fine-tuning, LLM fine-tuning, post-training optimization, prompt engineering, inference optimization, LLM agents, customizing LLMs, ensuring LLM alignment, Jailbreak defense, backdor defense, Watermark in LLMs, Model extraction and distillation, Adversarial Machine Learning, Generative Models, Self-supervised Learning, Federated Learning, Computer Vision, Natural Language Processing
Languages: (Proficient) Python; (Familiar) C, SQL
Developer Tools: VS Code, PyCharm, Conda, GitHub, Linux toolkits
Libraries / Frameworks: PyTorch, Huggingface, Transformers, OpenAI, scikit-learn, Keras, OpenCV, MySQL

Work Experience

Futurewei Technologies(Huawei US Research Institute) — 05/2024 – 08/2024
Research Intern (LLM advance solution) — Framingham, MA

- **LLM Analysis and Optimization:** Analyze output logits and attention scores to evaluate the impact of **LoRA fine-tuning**. Develop scripts to visualize and compare **model behavior** before and after fine-tuning.
- **LLM Knowledge Storage Locating:** Find that different expressions of the same facts activate similar layers in LLMs to **ensure alignment** between the fine-tuned LLM and the enterprise's private dataset.
- **Fine-tuning Resource Optimization:** Develop strategies for **building customized LLMs for enterprise clients**, focusing on minimizing resource usage while maintaining high performance. Implement solutions that reduce operational costs by 50% through efficient model fine-tuning techniques.

Selected Publications

[1] **Liang, J.,** Pang, R., Li, C., & Wang, T. (2024, July). Model extraction attacks revisited. *In Proceedings of the 19th ACM ASIA CCS - Asia Conference on Computer and Communications Security (pp. 1231-1245).*

[2] **Liang, J.,** Li, S., Cao, B., Jiang, W., & He, C. (2021). Omnilytics: A blockchain-based secure data market for decentralized machine learning. *International Workshop on Federated Learning for User Privacy and Data Confidentiality in Conjunction with ICML 2021*

[3] Jiang, T., Wang, Z., **Liang, J.,** Li, C., Wang, Y., & Wang, T. (2024). RobustKV: Defending large language models against jailbreak attacks via KV eviction. *The International Conference on Learning Representations (ICLR) 2025. url: https://arxiv.org/abs/2410.19937*

[4] 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 - IEEE Transactions on Pattern Analysis and Machine Intelligence.*

[5] Zhou, Q., Guo, S., Pan, J., **Liang, J.,** Xu, Z., & Zhou, J. (2023, June). PASS: patch automatic skip scheme for efficient real-time video perception on edge devices. *In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 3, pp. 3787-3795).*

[6] Li, J., Wei, G., **Liang, J.,** Ren, Y., Lee, P. P., & Zhang, X. (2022, May). Revisiting frequency analysis against encrypted deduplication via statistical distribution. *In IEEE INFOCOM 2022 - IEEE Conference on Computer Communications (pp. 290-299). IEEE.*

Works under review:

[7] **Liang, J.**, Wang, Y., Li, C., Zhu, R., Jiang, T., Gong, N., & Wang, T. (2025). GraphRAG under Fire. *Submitted to USENIX 2025. URL: https://arxiv.org/abs/2501.14050*

[8] **Liang, J.**, Wang, Z., Hong, L., Ji, S., & Wang, T. (2024). WaterPark: A robustness assessment of language model watermarking. *Submitted to ACL 2025. url: https://arxiv.org/abs/2411.13425*

[9] Liu, X.*, **Liang, J.***, Tang, L., You, C., Ye, M., & Xi, Z. (2024). Buckle up: Robustifying LLMs at every customization stage via data curation. *Submitted to ACL 2025. url: https://arxiv.org/abs/2410.02220*

[10] Li, C., **Liang, J.,** Cao, B., Chen, J., & Wang, T. (2024). Your agent can defend itself against backdoor attacks. *Submitted to ACL 2025.*

[11] 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. *Submitted to IEEE S&P 2025. url: https://arxiv.org/abs/2411.13144*

Academic Projects

Defensive Strategies for Large Language Models
08/2024 – Present

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

- **Backdoor Attack Defense for LLM Agents:** Developed a novel defense system that enables LLM agents to defend themselves against backdoor attacks by ensuring consistency between agent planning, execution, and user instructions.
- **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.
- **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.

Security Challenges in Large Language Models
10/2023 – Present

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

- **Evaluation Watermark's robustness and propose advanced attack method:** 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.
- **Potential Security Threats 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.
- **Using long CoT to jailbreak reasoning model:** Leverage the thinking process of the reasoning model to jailbreak the model itself.

Model Extraction Security Challenges on the Real MLaaS API
08/2022 – 10/2023

Research Assistant @ALPS-Lab (Dr. 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.
- **ME Threat Empirical Evaluation:** Used "MEBench" to study major MLaaS platforms (e.g., Amazon, Google), revealing significant vulnerabilities and inconsistent characteristics across tasks and platforms.
- **MLaaS Historical Analysis:** By analyzing data from 2020-2022, we identified key, evolving trends in ME vulnerabilities, highlighting the urgent need for stronger security in MLaaS platforms.

Video Perception Optimizing
10/2021 – 06/2022

Research Assistant @Pei-Lab (Dr. Song Guo), The Hong Kong Polytechnic University Hong Kong

- **Patch Automatic Skip (PASS) for Efficiency**: Engineered a plug-and-play module for CNN backbones, enabling patch-skippable architectures. PASS strategically skips specific image regions during training, optimizing computational efficiency while maintaining robustness.
- **Masked Anti-occlusion (MASK)**: Devised a cutting-edge semi-supervised learning framework that leverages semantic key-tokens to mask occluded image regions during training. It addresses visual occlusion challenges and enhances the model's reliability.

Against Encrypted Deduplication via Statistical Distribution
09/2020 – 09/2021

Research Assistant (Dr. Jingwei Li), University of Electronic Science and Technology of China Chengdu, China

- **Distribution-based Attack on Deduplication Storage System**: Conducted in-depth research into the vulnerabilities associated with encrypted deduplication storage systems. Developed a state-of-the-art distribution-based attack mechanism, using the inherent storage workload characteristics. This innovation significantly elevated inference precision.

Blockchain-based Secure Data Market for Federated Learning
03/2021 – 09/2021

Research Intern (Dr. Songze Li, Dr. Chaoyang He), The Hong Kong University of Science and Technology Hong Kong

- **Blockchain-based Data Market**: Architected and developed a blockchain-based platform geared towards the secure trade of machine learning data. Integrated advanced encryption and privacy-preserving mechanisms, ensuring data integrity, confidentiality, and traceable transactions in the marketplace.