Stuttgart, Germany
+49 0174 6064 083
zhixiong.zhuang96@gmail.com
zhixiong-zhuang
zhixiongzh.github.io
FHqC3vUAAAAJ&hl

Zhixiong Zhuang

Summary

Ph.D. candidate in AI security with 4 top-tier publications and 3 patents, funded by Bosch Research. Developed novel AI solutions, including LLM prompt leakage prevention, scalable image synthesis via generative priors for model extraction, and vulnerability analysis of multimodal LLMs. Currently contributing to RAG system evaluation at Bosch.

Education

CISPA Helmholtz Center for Information Security | Saarbrücken, Germany **Jan. 2023 - Jan. 2026**

- PhD. in Computer Science; Advisor: Prof. Dr. Mario Fritz and Dr. Maria-Irina Nicolae

Technical University of Munich | Munich, Germany **Oct. 2020 - Oct. 2022**

- MSc. in Robotics, Cognition and Intelligence; Advisor: Prof. Dr. Walter Stechele

Tongji University | Shanghai, China **Sept. 2015 - Jul. 2020**

- BSc. in Automotive Engineering (5-Year Program); Advisor: Prof. Dr. Guangqiang Wu

Industry Experience

Bosch Center for AI | Renningen, Germany | PhD Student **Jan. 2023 - Jan. 2026**

- Evaluated RAG systems to improve security and performance (on going).
- Protected LLM system prompts against extraction attacks, raising protection by 52% [1].
- Proposed victim-aware prompt optimization for image synthesis to improve proxy model accuracy by up to 22% [2].
- Proposed the first stealing attack on medical multimodal LLMs, achieving 92% of the victim model's performance. [3].
- Developed an environment-free stealing attack to replicate RL policies with up to 97% of the victim model's performance [4].

BMW | Munich, Germany | Master Thesis **Mar. 2022 - Aug. 2022**

- Proposed a genetic-algorithm based pruning method for panoramic segmentation on edge devices.

BMW | Munich, Germany | Intern **Aug. 2021 - Feb. 2022**

- Built a computer-vision-based system for battery cell leakage detection, enhancing battery safety.

Zhen Robotics | Beijing, China | Intern **Jul. 2020 - Oct. 2020**

- Developed and deployed 3D object detection for robots using OpenPCDet.

Continental AG | Shanghai, China | Intern **Mar. 2019 - Jul. 2019**

- Created standardized templates for technical solution proposals, streamlining new-project documentation and delivery efficiency.

Publications

[1] **Z. Zhuang**, M.-I. Nicolae, H.-P. Wang, M. Fritz. *ProxyPrompt: Securing System Prompts against Prompt Extraction Attacks*. **NeurIPS** 2025 (under review); 3 patents (submitted) [PDF]

[2] **Z. Zhuang**, H.-P. Wang, M.-I. Nicolae, M. Fritz. *Stealix: Model Stealing via Prompt Evolution*. **ICML** 2025. [PDF]

[3] Y. Shen*, **Z. Zhuang*** (co-first), K. Yuan, M.-I. Nicolae, N. Navab, N. Padoy, M. Fritz. *Medical Multimodal Model Stealing Attacks via Adversarial Domain Alignment*. **AAAI** 2025 (**Oral, top 4.6%**). [PDF]

[4] **Z. Zhuang**, M.-I. Nicolae, M. Fritz. *Stealthy Imitation: Reward-guided Environment-free Policy Stealing*. **ICML** 2024. [PDF]

Skills

- **LLM Expertise:** Prompt optimization, Multimodal LLM fine-tuning, LLM-as-judge
- **ML Frameworks & Tools:** PyTorch, TensorFlow
- **Programming Languages:** Python, C++
- **Languages:** Chinese (Native), English (Fluent), German (Intermediate)

Achievements

- **Reviewer:** NeurIPS 2025, AISec 2025
- **Mentoring:** Yaling Shen (MSc TUM; currently PhD candidate at Monash University)