(+1) 540-934-8165
tranhn137@gmail.com
tranhn137.github.io
Google Scholar

Tran Huynh Ph.D. Student

Education

Department of Electrical & Computer Engineering, Virginia Tech — Virginia, USA
Ph.D. in Computer Engineering — *2024 - 2029 (expected)*

Advisor: Prof. Ruoxi Jia

University of Information Technology (UIT) — Ho Chi Minh, Vietnam
Bachelor of Computer Science — *2018 - 2022*

Faculty top 1 scorer of the university entrance exam

CPA: 8.53/10.0, Very Good Degree

Research Experience

VinAI Research — 03/2022 - 06/2024
Research Resident

Project: ***"COMBAT: Alternated Training for Effective Clean-Label Backdoor Attacks"***
This project aims to enhance the efficiency of clean-label backdoor attacks by creating a training approach that involves the simultaneous training of a trigger generator and the poisoning of a surrogate model. The suggested attack attains a near-perfect success rate and exhibits great adaptability to meet diverse constraints, such as visual imperceptibility, input awareness, and multi-target attack objectives.

Project: ***"Data Poisoning Quantization Backdoor Attack"***
This project explores the potential risk of data poisoning backdoor attack in model quantization. While previous attacks on quantization model requires full model training control, this method aims to achieve the attacker's goals based on data poisoning only with zero prior knowledge of the target victim model.

Department of Electrical & Computer Engineering, Virginia Tech — 08/2024 - Present
Ph.D. Student in Computer Engineering

Project: ***"BEEAR: Embedding-based Adversarial Removal of Safety Backdoors in Instruction-tuned Language Models"***
This project addresses the challenge of safety backdoor attacks in large language models (LLMs). Unlike previous approaches that focus on specific trigger mechanisms or attack types, this method leverages the observation that backdoor triggers induce a uniform drift in the model's embedding space and utilizes bi-level optimization to defend against these triggers effectively, without needing prior knowledge of the attack specifics or trigger location.

Project: ***"Inference-Time Personalized Safety Control via Paired Difference-in-Means Intervention"***
This project introduces a training-free framework for personalized LLM safety, enabling users to suppress specific undesired content (e.g., violence, political ideology) without costly retraining. Unlike standard universal alignment, this method employs Paired Contrast Mean Shift (PCMS) to estimate user-specific harm directions by isolating activation differences in topic-matched prompt pairs. By applying these interventions at inference time using a quantile-based activation threshold, the system effectively reduces harmful content intensity while preserving general model utility and reasoning capabilities, as validated across multiple open-weight models and safety benchmarks.

Project: ***"Adversarial Déjà Vu: Jailbreak Dictionary Learning for Stronger Generalization to Unseen Attacks"***
This project addresses the critical challenge of defending LLMs against unseen jailbreak attacks by proposing the Adversarial Déjà Vu hypothesis: novel attacks are largely sparse recombinations of existing adversarial skills rather than fundamentally new strategies. To validate this, the work introduces a large-scale automated pipeline to extract and compress attack techniques from two years of literature into a sparse dictionary of primitives. Guided by this insight, the project develops Adversarial Skill Compositional Training (ASCoT), a defense method that trains models on diverse compositions of these skill primitives. Empirical results demonstrate that ASCoT significantly improves robustness against unseen attacks, including complex multi-turn jailbreaks, while maintaining low over-refusal rates on benign queries.

Amazon Nova AI Challenge 11/2024 - 06/2025
Selected Participant

Project: ***"Data is all you need (almost): Iterative Synthetic Instruction Tuning for Secure Code Generation"***
This project addresses the scarcity of high-quality secure code data by designing an iterative synthetic data generation pipeline that leverages iterative critique and refinement. This approach significantly improved the model's ability to generate secure code and follow complex security constraints, demonstrating the efficacy of synthetic instruction tuning for domain-specific alignment without relying on extensive human-annotated datasets.

Publications

1. **Tran Huynh**, Dang Nguyen, Tung Pham, Anh Tran. COMBAT: Alternated Training for Effective Clean-Label Backdoor Attacks. In proceedings of *Association for the Advancement of Artificial Intelligence (AAAI)*, 2024.
2. **Tran Huynh**, Anh Tran, Khoa Doan, Tung Pham. Data Poisoning Quantization Backdoor Attack. In proceedings of *European Conference on Computer Vision (ECCV)*, 2024.
3. Yi Zeng, Weiyu Sun, **Tran Huynh**, Dawn Song, Bo Li, Ruoxi Jia. BEEAR: Embedding-based Adversarial Removal of Safety Backdoors in Instruction-tuned Language Models. In proceedings of *Empirical Methods in Natural Language Processing (EMNLP)*, 2024.
4. **Tran Huynh**, Ruoxi Jia. Inference-Time Personalized Safety Control via Paired Difference-in-Means Intervention. In proceedings of *International Conference on Learning Representations (ICLR)*, 2026.
5. Mahavir Dabas, **Tran Huynh**, Nikhil Reddy Billa, Jiachen T. Wang, Peng Gao, Charith Peris, Yao Ma, Rahul Gupta, Ming Jin, Prateek Mittal, Ruoxi Jia. Adversarial Déjà Vu: Jailbreak Dictionary Learning for Stronger Generalization to Unseen Attacks. In proceedings of *International Conference on Learning Representations (ICLR)*, 2026.

Awards and Honors

- Selected Participant, Amazon Nova AI Challenge 2025: Recognized among top teams to develop innovative AI solutions using Amazon's Nova foundation models.
- Certificate of Attendance in the National Mathematics Olympiad for Students in 2019
- First prize in "UIT's Leader" contest in 2019
- Scholarship for top scorers in the university entrance exam of University of Information Technology (UIT) in 2018
- Scholarships for outstanding academic achievement in 6 academic semesters at UIT
- Silver medal in the National Olympic Smart English in 2017

Skills

Programming languages: Python, C/C++

ML Frameworks: PyTorch, TensorFlow

Libraries & Tools: NumPy, Pandas, Git, Docker

Academic Services

Reviewer for: CVPR 2023, NeurIPS 2023, ICLR 2024, CVPR 2024, ECCV 2024, AAAI 2025, Neurips 2025, ICLR 2026.