ZIYU LIU

+1 (814) 699-2504 | zliu48@binghamton.edu | LinkedIn | Homepage

Research Interests

Efficient and trustworthy AI systems, including compression, distillation, security, and on-device deployment of language models.

Education

Binghamton University Expected graduation: 12/27
Ph.D. in Electrical and Computer Engineering

Northeastern University 05/22
M.S. in Electrical Engineering

Pennsylvania State University 05/20
B.S. in Electrical Engineering

Research Experience

CORE: Protecting LLMs Against Functionality Stealing *Under Review*
Binghamton University

- Identified **critical MLPs** by removing weights associated with outlier activations from individual layers and evaluating the impact on LLM performance.
- Developed a defense using fixed encodings and a learned recovery module to hinder reconstruction of protected components by occasionally modifying observable outputs.
- Analyzed how these output changes can make it harder to train an accurate surrogate using MSE loss.
- Evaluated on Llama-2-7B, Llama-3-8B, and Qwen3-8B, retaining **96.0–98.8%** of original average zero-shot accuracy, while surrogate replacements scored **26.0–47.3%** lower than the protected models.

Scout: Query-Efficient Extraction of LLM Specialists with Undisclosed Tasks *Under Review*
Binghamton University

- Developed a query-efficient method to extract LLM specialists with undisclosed tasks by comparing response likelihoods under the victim and its public base model.
- Built an extraction pipeline that groups public prompts into semantic regions, uses response scores to allocate queries with Thompson sampling, and fine-tunes surrogate models with LoRA.
- Achieved a **6.1–11.3%** relative improvement in GSM8K accuracy over three query-selection baselines under the same 1,500-query budget when extracting a Llama-3.1-8B math specialist.

TBNet: Efficient Model Protection through Neural Architecture Optimization **DAC 2024**
Northeastern University

- Designed a two-branch neural network that runs a partial model in the untrusted environment to reduce latency, while keeping a lightweight protected branch inside the trusted execution environment (TEE).
- Developed joint training and iterative channel pruning to reduce secure-branch memory and computation.
- Deployed and evaluated on Raspberry Pi 3 with OP-TEE, achieving up to **2.45**× reduction in secure memory usage and **1.22**× inference speedup over full-model execution inside the TEE.

Securing PCIe Communication in Multi-tenant Cloud FPGAs **TODAES 2024**
Northeastern University

- Investigated how voltage-based fault injection through a shared power distribution network can compromise PCIe data transmission in multi-tenant FPGAs.
- Built a Virtex-6 FPGA prototype and characterized voltage fluctuations and fault behavior across attack patterns using non-invasive electromagnetic (EM) measurements.
- Developed a frequency-scaling defense that prevented fault injection during PCIe data transmission with **10.13%** data-throughput overhead.

MirrorNet: Protecting On-Device Models with a Lightweight Secure Branch **ICCAD 2023**
Northeastern University

- Designed a neural network with a lightweight branch inside a TEE, enabling accurate inference while limiting the accuracy of the exposed model available to attackers.
- Built a configuration generator and a latency predictor based on hardware profiling and regression to select secure-branch designs that balance model protection and inference overhead.
- Achieved **91.9%** accuracy versus **73.3%** for the exposed backbone on VGG-7/CIFAR-10 with only 0.99% inference latency overhead.

Publications

TBNet: A Neural Architectural Defense Framework Facilitating DNN Model Protection in Trusted Execution Environments.
Ziyu Liu, Tong Zhou, Yukui Luo, Xiaolin Xu.
ACM/IEEE Design Automation Conference (**DAC**), 2024.

Watch Out for the Inherent Vulnerabilities in Developing Multi-tenant Cloud-FPGA: Communication Protocols.
Ziyu Liu, Yukui Luo, Yuheng Zhang, Shijin Duan, Xiaolin Xu.
ACM Transactions on Design Automation of Electronic Systems (**TODAES**), 2024.

MirrorNet: A TEE-Friendly Framework for Secure On-device DNN Inference.
Ziyu Liu, Yukui Luo, Shijin Duan, Tong Zhou, Xiaolin Xu.
IEEE/ACM International Conference on Computer-Aided Design (**ICCAD**), 2023.

Industry Experience

Software Engineer Intern 07/21 – 01/22
Philips Lifeline, Framingham, MA

- Developed a C# Windows application to configure Inovonics Repeater network IDs in senior living facilities.
- Designed a file management system for RTL8722DM MINI boards.
- Evaluated fall-detection accuracy using wearable-device simulations to assess system reliability.

Technical Skills

Programming: Python, C/C++, MATLAB, SQL.
ML Frameworks & Tools: PyTorch, Hugging Face Transformers, vLLM, lm-evaluation-harness.
ML Techniques: supervised fine-tuning (SFT), LoRA, knowledge distillation, model compression/pruning.
Systems & Hardware: Arm TrustZone (OP-TEE), Raspberry Pi, Linux, Xilinx FPGAs, Verilog/SystemVerilog.

Teaching Experience

Instructor *Binghamton University*
EECE 251: Digital Logic Design Summer 2025

Teaching Assistant *Northeastern University*
EECE 2311: Intro to Digital Design and Computer Architecture Fall 2024

Teaching Assistant *Northeastern University*
EECE 2322/2323: Fundamentals of Digital Design and Computer Organization Fall 2023, Spring 2024

Service

Volunteer, New England Hardware Security Workshop (NEHWS), 2023.
Volunteer, Cyber Physical AI Trust and Integrity (CAITI) cluster, 2026.