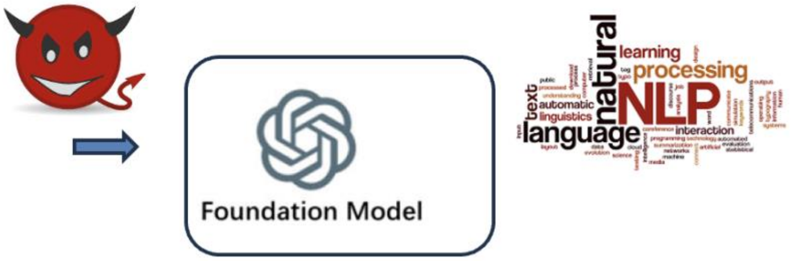
THE UNIVERSITY OF ALABAMA | College of Engineering Computer Science

A list of ongoing *AI/ML and Security/Privacy* projects in Trustworthy AI Lab is below.
If interested to involve, please contact Dr. Sayanton Dibbo (sdibbo@ua.edu)

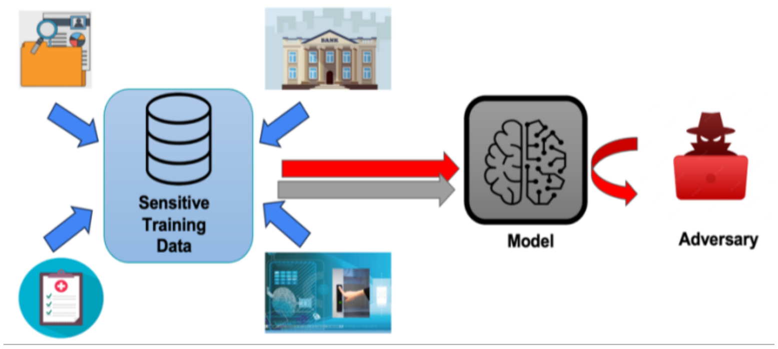


Project #1: Securing Chain-of-Thought LLMs Against Poisoning Adversarial Attacks

- Vulnerabilities in chain-of-thought (CoT) reasoning large language models (LLMs)
- Under what characteristics are these CoTs are more vulnerable?
- What mitigation techniques can be effective?

Project #2: Flipping the Binary: Adversarial Vulnerability of AI with Partial Class Knowledge

- Tabular data adversarial attacks
- The attacker has limited knowledge
- Study the adversarial vulnerability of binary classifiers and transferability

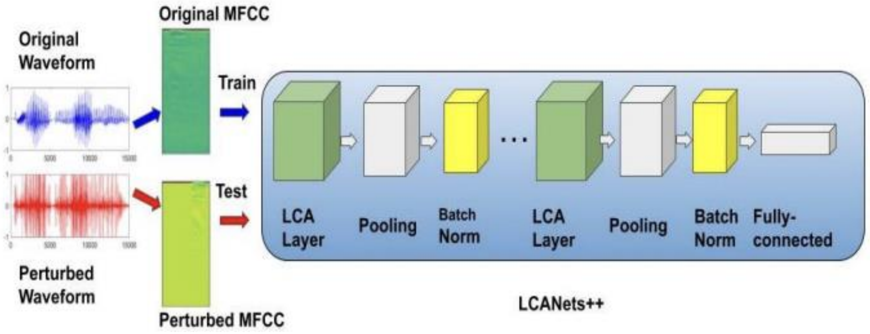


Project #3: Privacy Through Forgetting:
Exploring Machine Unlearning in Multimodal Foundational Models (Text, Audio, and Image)

- Literature review of machine unlearning techniques to enhance privacy
- Multimodal AI models vulnerabilities
- Study effectiveness and challenges in multimodal privacy attack settings

Project #4: Who's in the Data? Membership Inference Attacks on RL Systems in the Audio Domain

- Analyze privacy vulnerability in the audio domain
- Investigate privacy attack risks against RL models
- Propose defenses against privacy attacks


Original MFCC
Original Waveform
Train
Test
Perturbed Waveform
Perturbed MFCC
LCA Layer
Pooling
Batch Norm
LCA Layer
Pooling
Batch Norm
Fully-connected
LCANets++