Project Summary

Overview.
Large language model (LLM) services are increasingly deployed as black-box APIs across enterprise, government, and research settings. Model extraction attacks, in which adversaries systematically query a target model to replicate its behavior, pose a direct threat to the intellectual property and operational integrity of these services. Existing defenses are predominantly reactive, relying on watermarking that offer little practical protection, particularly against state-level adversaries. This project develops *Trap, Intervene, and Corrupt* (TIC), a proactive three-layer defense framework against LLM model extraction. Thrust 1 builds extraction-aware sensing mechanisms using knowledge honeypots and latent-space query trajectory analysis to identify attackers, including those operating across distributed accounts. Thrust 2 develops active intervention strategies that trap suspicious queries within semantically dense knowledge regions and serve highly variable output distributions to detected adversaries, degrading extraction efficiency without affecting legitimate users. Thrust 3 constructs retaliatory mechanisms that flood trapped attackers with semantically ambiguous knowledge and inject rationale-corrupted training signals, causing silent model degradation that resists forensic detection. In parallel, the project develops an interactive educational platform for LLM security, extending existing course infrastructure used in FSU CIS 4930/CAP 5638 and CIS 4930/CAP 5771, with outreach workshops co-organized with Florida A&M University and Tallahassee State College.

Intellectual Merit.
This project advances knowledge in adversarial machine learning, LLM security, and AI-powered defense systems. Four contributions are central. First, the project introduces a honeypot-based extraction sensing paradigm grounded in knowledge graph construction, enabling lightweight, distribution-agnostic detection of coordinated model extraction without access to attacker identity. Second, it develops a latent-space behavioral fingerprinting method for query trajectory analysis, producing a generalizable anomaly detection technique applicable beyond model extraction to other query-level threat scenarios. Third, the active intervention thrust contributes novel output distribution randomization strategies that provably increase attacker sample complexity while preserving response quality for benign users, advancing the theory of query-adaptive defense. Fourth, the rationale corruption mechanism introduces a new threat surface and defense primitive: injecting structurally flawed but outcome-correct reasoning chains into attacker training data, advancing understanding of how chain-of-thought supervision can be weaponized and defended. Together these thrusts produce a unified, principled framework that addresses the fundamental tension between service utility and model confidentiality in LLM deployment.

Broader Impacts.
This project delivers impact across four dimensions. First, the TIC framework addresses a critical vulnerability in commercial and government LLM deployments by providing the first end-to-end proactive extraction defense, with direct applications to API-based AI services. All code, datasets, and benchmarks will be released under open licenses, enabling adoption and extension by the broader research community. Second, the educational platform builds on the PI's existing course infrastructure and provides browser-accessible, interactive modules on LLM security, adversarial attacks, and extraction defenses, designed for deployment without specialized hardware. These materials will be integrated into FSU graduate and undergraduate AI courses and disseminated through co-organized workshops and seminars with FAMU and TSC. Third, the project trains the next generation of AI security researchers through hands-on red teaming exercises, capstone attack-defense projects, and practitioner-facing workshops, directly expanding the national pipeline of experts with combined AI and cybersecurity competencies. Fourth, the independent evaluation component, conducted through external assessment partnerships, will produce validated measurement instruments for LLM security education that serve the broader AI security education community.

Keywords: LLM Security; Model Extraction Defense; Adversarial Machine Learning; AI Education.