ACM CCS

Overall Review

ACM Conference on Computer and Communications Security: Economic Incentives, Human Factors, and System Integration (2022-2025)

Research at the ACM Conference on Computer and Communications Security (CCS) from 2022 through 2025 directly confronts what has been characterized as a "significant, practical gap" between theoretical adversarial machine learning research and the security challenges faced by deployed systems (Hammoudeh & Lowd, 2022). Unlike work focusing primarily on attack sophistication or isolated defensive mechanisms, ACM CCS contributions reveal how economic incentives, human perception, and system integration fundamentally shape both the threat landscape and viable defense strategies. This research demonstrates that bridging the theory-practice gap requires understanding not just technical vulnerabilities but also the business logic, usability constraints, and architectural decisions that govern real-world machine learning deployments.

Economic Incentives and Intellectual Property Vulnerabilities

The economics of adversarial machine learning fundamentally differs from academic threat modeling. Commercial machine learning services create powerful financial incentives for intellectual property theft that render many protective mechanisms inadequate. Stealing a pre-trained Self-Supervised Learning encoder from Encoder-as-a-Service platforms costs far less than training one from scratch, creating strong economic motivation for model extraction attacks against commercial services (Cong et al., 2022). This vulnerability drives defensive

countermeasures like SSLGuard watermarking to protect intellectual property, though such protections prove fragile under realistic attack scenarios.

The inadequacy of intellectual property defenses becomes apparent when attacks operate under minimal resource constraints. Stealing proprietary decoding algorithms and hyperparameters from Large Language Model APIs via black-box access proves remarkably inexpensive despite the high commercial value of this intellectual property (Naseh et al., 2023). More severely, Neural Dehydration attacks effectively remove embedded watermarks from deep neural networks using less than 2% of the original training data, or even in completely data-free settings, demonstrating that current watermarking schemes fail when organizations cannot guarantee strict data isolation (Lu et al., 2024). This economic asymmetry—where attacks cost orders of magnitude less than defenses or the assets they protect—creates fundamental challenges for commercial deployment.

Defensive efforts to protect high-value commercial assets must therefore balance robustness against operational requirements. Watermarking techniques for Large Language Model-based code generation APIs, such as ToSyn, must withstand attacks from fully knowledgeable adversaries while minimizing disruption to API functionality and maintaining acceptable throughput for commercial users (Li et al., 2023). Similarly, defenses against model extraction like Beowulf reshape decision regions using synthetic dummy classes and noise to make models difficult to replicate via API queries, though such approaches introduce complexity that may affect service quality (Gong et al., 2024). The tension between strong protection and commercial viability remains an open challenge.

The Disconnect Between Mathematical Metrics and Human Perception

A critical gap in adversarial machine learning research stems from reliance on mathematical perturbation metrics that fail to capture actual human perception. Conventional distance metrics like Lp norms—measuring pixel-level differences—do not reliably predict whether humans perceive adversarial perturbations as anomalous or suspicious. The Perception-Aware Attack framework addresses this disconnect by validating attacks against commercial systems like YouTube's copyright detector, demonstrating that attacks optimized for mathematical

imperceptibility may still be readily apparent to human observers (Duan et al., 2022).

This perception gap proves particularly severe in safety-critical applications involving human operators. Research on autonomous driving reveals that widely adopted digital measures of attack inconspicuousness fail to predict whether human drivers actually notice physical adversarial attacks on traffic signs (Ma et al., 2024). Using virtual reality and eye-tracking systems, the Avara framework demonstrates that Lp norms and other common metrics do not adequately address real-world safety risks in human-in-the-loop systems. An attack deemed "imperceptible" by mathematical standards may be immediately obvious to a driver, while conversely, attacks that violate mathematical constraints might go unnoticed in realistic driving conditions. This mismatch between academic evaluation and practical safety requirements suggests that current research may fundamentally mischaracterize risks in deployed systems.

Physical constraints further complicate the perception challenge. AttrackZone leverages 3D point cloud data to identify physically realizable "attack zones" where projector-based adversarial manipulations can successfully compromise object trackers while remaining within environmental constraints (Muller et al., 2022). Such physically grounded approaches recognize that mathematical optimization must be constrained by the actual physics of attack deployment—light propagation, viewing angles, environmental interference—rather than operating in abstract feature spaces.

Minimal-Information and Zero-Knowledge Attack Feasibility

Real-world adversaries rarely possess the comprehensive system knowledge assumed in academic threat models. Recognition of this gap has driven development of attacks succeeding under minimal-information constraints that more accurately reflect practical scenarios. Traditional clean-label backdoor attacks require adversaries to possess full knowledge of victim model training data—a condition rarely met in real-world contexts like crowdsourced data collection. The Narcissus attack addresses this limitation by demonstrating effective clean-label backdoors using only limited information about the training process (Zeng et al., 2023).

Similarly, attacks against machine learning-based Android malware detection historically relied on detailed knowledge of feature spaces, model parameters, or training sets. AdvDroidZero succeeds under a more pragmatic zero-knowledge setting where attackers lack these detailed specifications (Li et al., 2023). This shift toward zero-knowledge attacks reveals that defensive strategies predicated on confidentiality of implementation details may provide inadequate protection. Even more severely, PLeak automatically optimizes zero-query adversarial prompts to steal confidential system prompts from Large Language Model applications, offering a scalable attack vector against commercial services that far surpasses manually crafted methods (Hui et al., 2024).

The minimal-information paradigm extends to graph-structured data. MNEMON demonstrates model-agnostic graph recovery attacks assuming only access to node embeddings—a realistic scenario in enterprise environments like vertical federated learning where direct model interaction is prohibited by architectural constraints (Shen et al., 2022). These attacks reveal that even carefully architected systems designed to limit information exposure remain vulnerable when adversaries can infer sensitive structure from limited observations.

Recent work further demonstrates that attackers can achieve sophisticated objectives with remarkably little knowledge. The UnivIntruder framework creates universal, transferable, and targeted adversarial examples against black-box systems, achieving high success rates on commercial services like Google Search and GPT-4 without requiring high query counts or access to target training data (Xu et al., 2025). This universality and transferability suggests that defenses relying on query limiting or model confidentiality may prove ineffective against practical adversaries.

Deployment Barriers in Production Environments

The gap between theoretical attacks and practical deployment manifests acutely when targeting real-world systems. Standard adversarial example generation techniques succeed against only 6.53% of real-world deep neural network models deployed in Android applications due to constraints like proprietary model formats, encryption, and access restrictions (Deng et al., 2022). Overcoming these barriers requires model extraction and interface reasoning—increasing attack complexity but achieving 47.35% success rates against deployed systems.

This stark difference between attacking research models and production deployments underscores how implementation details create practical security barriers absent from academic evaluations.

Architectural decisions intended to improve performance or protect privacy often inadvertently create new vulnerabilities. Multi-Exit Networks, designed to improve inference efficiency by allowing early termination for simple inputs, leak membership information through their exit depth configuration—an unintended side channel arising from the optimization-driven architecture (Li et al., 2022). Similarly, incorporating Large Language Models into split learning frameworks creates significant data leakage risks by exploiting the "Not-too-far" property of fine-tuning to perform successful data reconstruction despite the privacy guarantees split learning purports to provide (Chen et al., 2024).

Efficiency-driven optimizations create additional attack surfaces. Research demonstrates that poisoning training data can dramatically reduce the computational cost of privacy attacks by eliminating expensive steps like training shadow models (Tramer et al., 2022). This cost reduction makes previously impractical attacks viable for resource-constrained adversaries, fundamentally altering the threat landscape. The finding that protective measures themselves can be compromised through data poisoning reveals how defenses optimized for one threat model may create vulnerabilities to adaptive adversaries.

System-Level Integration and Weakest-Link Vulnerabilities

As machine learning components integrate into complex production systems, individual model robustness proves insufficient for system-level security. A single non-robust component can compromise an entire security pipeline regardless of how well other components are hardened. The attack against Magika—Google's file-type classifier—demonstrates this principle by exploiting a publicly known machine learning component to compromise Gmail's entire malware detection pipeline (Nasr et al., 2025). Even though other pipeline components may be robust, the weakest link determines overall system security.

This weakest-link problem extends to training pipelines and model supply chains. Attacks targeting Vision-Language Models inject adversarial mislabeled poison into training data for Text-to-Image models, corrupting outputs at generation time

despite the models appearing to train normally (Wu et al., 2025). Knowledge distillation amplifies such vulnerabilities: the BIASED-ROOTS attack demonstrates how malicious biases can cascade from teacher to student language models, propagating adversarial influence throughout model lineages (Chaudhari et al., 2025). These supply chain attacks reveal that security must be maintained throughout the entire model development lifecycle, not just at deployment.

Physical infrastructure integration introduces additional system-level vulnerabilities. Attacks on autonomous driving systems exploit weaknesses in visual perception and online mapping pipelines through low-cost physical-world interference, demonstrating that safety-critical systems remain vulnerable where digital defenses meet physical reality (Ma et al., 2025; Lou et al., 2025). ControlLoc demonstrates physical-world hijacking of visual perception in autonomous vehicles, while research on asymmetry vulnerabilities exposes fundamental model biases that physical attacks can exploit in mapping systems essential for navigation.

Computational Barriers to Formal Verification

Fundamental computational complexity limits the applicability of rigorous security guarantees. Verifying the robustness of general decision tree ensembles against adversarial threats proves NP-hard, creating an insurmountable barrier for providing formal guarantees on complex models (Calzavara et al., 2023). This computational infeasibility motivated development of "verifiable learning" techniques that constrain model complexity to allow polynomial-time verification—accepting reduced model capacity as the cost of formal guarantees. This trade-off exemplifies how theoretical rigor conflicts with practical deployment needs.

The verification challenge extends beyond computational complexity to defensive circumvention. Stateful Defense Models designed specifically to thwart query-based black-box attacks by tracking similar queries prove highly vulnerable to adaptive strategies that exploit information inadvertently leaked by the defense mechanism itself (Qin et al., 2023). More concerning, models enforced with security properties—designed to provide verifiable guarantees—demonstrate significantly higher susceptibility to model stealing attacks than traditionally trained models. This reveals a fundamental tension: enhancing one security property often degrades protection against orthogonal threats.

Adaptive Defenses for Realistic Threat Models

Recognition that standard, one-size-fits-all defenses fail in practical deployments has driven development of adaptive, context-aware defensive strategies. VisionGuard successfully detects physical adversarial attacks against autonomous driving perception modules by exploiting their inherent spatiotemporal inconsistency across video frames, leveraging reliable internal kinetic data from GPS and IMU sensors rather than relying on complex digital certification or visual features susceptible to adversarial manipulation (Han et al., 2024). This approach demonstrates how defenses can leverage domain-specific physical constraints and sensor fusion rather than purely digital robustness mechanisms.

For distributed learning environments, personalized defenses address the failure of generalized approaches when deployed across heterogeneous data distributions and resource-constrained devices. FilterFL and Sylva develop personalized adversarial robustness solutions tailored for the specific constraints of federated learning and edge computing, recognizing that defenses must adapt to local data characteristics and computational budgets rather than assuming uniform deployment conditions (Yang et al., 2025; Qi et al., 2025). These personalized approaches accept that different participants in distributed systems require different defensive strategies based on their specific risk profiles and resource availability.

Low computational cost proves critical for defense adoption. Multi-domain Trojan detection methods aim for broad applicability across data domains while remaining computationally inexpensive, contrasting with prior approaches requiring specialized resources for model retraining or trigger reconstruction (Rajabi et al., 2023). Similarly, replacing subjective, heuristic-based malware detection methods with robust, learning-based models designed to identify sophisticated and previously unseen packed binaries addresses the continuous evolution of adversarial software while maintaining acceptable performance (Li et al., 2025).

Evolving Threats Against Emerging Technologies

The rapid deployment of generative AI and large language models creates urgent security challenges where defenses lag substantially behind evolving threats. Research tracking real-world jailbreak attempts via platforms like JailbreakHub

confirms that Large Language Model safety safeguards including Reinforcement Learning from Human Feedback frequently fail against continuously evolving, in-the-wild prompt injection and obfuscation strategies (Shen et al., 2024). This empirical evidence demonstrates that standard defenses cannot keep pace with adversarial innovation in practical deployment.

Public safety concerns associated with generative models demand targeted defensive strategies. Research documenting widespread misuse of text-to-image models to generate unsafe images and hateful memes underscores the urgent need for safeguard tools that address scalable, real-world threats rather than theoretical worst-case scenarios (Qu et al., 2023). However, defensive techniques themselves face adversarial evasion: watermark-based detection of AI-generated content proves vulnerable to removal attacks, limiting effectiveness of provenance tracking mechanisms (Jiang et al., 2023).

Emerging architectures introduce novel attack surfaces. SuperNets—neural architectures supporting multiple sub-networks with different accuracy-latency trade-offs—enable precisely controlled backdoor activation based on which sub-network operates during inference, as demonstrated by the VillainNet attack (Anonymous, 2025). This architectural complexity creates stealth opportunities where backdoors activate only under specific operational conditions, evading defenses designed to detect universal triggers. As model architectures grow more complex to meet diverse deployment requirements, the attack surface expands in ways not addressed by current defensive strategies.

Privacy-Security Trade-offs in Defensive Mechanisms

Efforts to enhance security properties often create or exacerbate privacy vulnerabilities. Models trained with robustness enhancements designed to resist adversarial examples demonstrate increased susceptibility to model stealing attacks compared to traditionally trained models (Qin et al., 2023). This privacy-security trade-off reveals that defenses cannot be evaluated in isolation—improving resistance to one threat may create new vulnerabilities to others.

The trade-off extends to privacy-preserving training mechanisms. Holistic evaluation frameworks for Differentially Private Machine Learning reveal complex interactions between privacy guarantees, model utility, and vulnerability to various attack types (Wei et al., 2023). Achieving meaningful privacy protection often

requires privacy budget settings that substantially degrade model accuracy, creating practical barriers to deployment. Moreover, adaptive adversaries can exploit defensive mechanisms: poisoning attacks specifically designed to boost information leakage demonstrate that Differential Privacy alone provides inadequate protection when training data can be maliciously manipulated (Tramer et al., 2022).

References

Anonymous. (2025). VillainNet: Targeted Poisoning Attacks Against SuperNets Along the Accuracy-Latency Pareto Frontier. *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*.

Calzavara, S., Cazzaro, L., Pibiri, G. E., & Prezza, N. (2023). Verifiable Learning for Robust Tree Ensembles. *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*.

Chaudhari, H., Hayes, J., Jagielski, M., Shumailov, I., Nasr, M., & Oprea, A. (2025). Cascading Adversarial Bias from Injection to Distillation in Language Models. *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*.

Chen, G., Qin, Z., Yang, M., Zhou, Y., Fan, T., Du, T., & Xu, Z. (2024). Unveiling the Vulnerability of Private Fine-Tuning in Split-Based Frameworks for Large Language Models: A Bidirectionally Enhanced Attack. *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*.

Cong, T., He, X., & Zhang, Y. (2022). SSLGuard: A Watermarking Scheme for Self-supervised Learning Pre-trained Encoders. *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*.

Deng, Z., Chen, K., Meng, G., Zhang, X., Xu, K., & Cheng, Y. (2022). Understanding Real-world Threats to Deep Learning Models in Android Apps. *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*.

Duan, R., Qu, Z., Zhao, S., Ding, L., Liu, Y., & Lu, Z. (2022). Perception-Aware Attack: Creating Adversarial Music via Reverse-Engineering Human Perception. *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*.

Gong, X., Wei, R., Wang, Z., Sun, Y., Peng, J., Chen, Y., & Wang, Q. (2024). Beowulf: Mitigating Model Extraction Attacks Via Reshaping Decision Regions. *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*.

Hammoudeh, Z., & Lowd, D. (2022). Training Data Inference with Restricted Sample Access. *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*.

Han, X., Wang, H., Zhao, K., Deng, G., Xu, Y., Liu, H., Qiu, H., & Zhang, T. (2024). VisionGuard: Secure and Robust Visual Perception of Autonomous Vehicles in Practice. *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*.

Hui, B., Yuan, H., Gong, N., Burlina, P., & Cao, Y. (2024). PLeak: Prompt Leaking Attacks against Large Language Model Applications. *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*.

Jiang, Z., Zhang, J., & Gong, N. Z. (2023). Evading Watermark-based Detection of AI-Generated Content. *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*.

Li, P., Xia, Y., Zhang, X., & Ji, S. (2023). Efficient Query-Based Attack against ML-Based Android Malware Detection under Zero Knowledge Setting. *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*.

Li, S., Ming, J., Liu, L., Yang, L., Zhang, N., & Jia, C. (2025). Adversarially Robust Assembly Language Model for Packed Executables Detection. *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*.

Li, Z., Liu, Y., He, X., Yu, N., Backes, M., & Zhang, Y. (2022). Membership Inference Attacks by Exploiting Loss Trajectory. *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*.

Li, Z., Wang, C., Wang, S., & Gao, C. (2023). Protecting Intellectual Property of Large Language Model-Based Code Generation APIs via Watermarks. *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*.

Lou, Y., Hu, H., Song, Q., Xu, Q., Zhu, Y., Tan, R., Lee, W.-B., & Wang, J. (2025). Asymmetry Vulnerability and Physical Attacks on Online Map Construction for

Autonomous Driving. *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*.

Lu, Y., Li, W., Zhang, M., Pan, X., & Yang, M. (2024). Neural Dehydration: Effective Erasure of Black-box Watermarks from DNNs with Limited Data. *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*.

Ma, C., Wang, N., Zhao, Z., Wang, Q., Chen, Q. A., & Shen, C. (2025). ControlLoc: Physical-World Hijacking Attack on Visual Perception in Autonomous Driving. *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*.

Ma, X., Zhang, C., Zhu, H., Camp, L. J., Li, M., & Liao, X. (2024). Avara: A Uniform Evaluation System for Perceptibility Analysis Against Adversarial Object Evasion Attacks. *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*.

Muller, R., Man, Y., Celik, Z. B., Li, M., & Gerdes, R. (2022). AttrackZone: An Automated Black-box Attack Generation System for Autonomous Vehicles. *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*.

Naseh, A., Krishna, K., Iyyer, M., & Houmansadr, A. (2023). Stealing the Decoding Algorithms of Large Language Models. *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*.

Nasr, M., Fratantonio, Y., Invernizzi, L., Albertini, A., Farah, L., Petit-Bianco, A., Terzis, A., Thomas, K., Bursztein, E., & Carlini, N. (2025). Evaluating the Robustness of a Production Malware Detection System to Transferable Adversarial Attacks. *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*.

Qi, T., Xue, L., Zhan, Y., & Ma, X. (2025). Sylva: Tailoring Personalized Adversarial Defense in Pre-trained Models via Collaborative Fine-tuning. *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*.

Qin, Y., Fu, Z., Deng, C., Liao, X., Zhang, J., & Duan, H. (2023). Stolen Risks of Models with Security Properties. *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*.

Qu, Y., Shen, X., He, X., Backes, M., Zannettou, S., & Zhang, Y. (2023). Unsafe Diffusion: On the Generation of Unsafe Images and Hateful Memes From Text-To-

Image Models. *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*.

Rajabi, A., Asokraj, S., Jiang, F., Niu, L., Ramasubramanian, B., Ritcey, J., & Poovendran, R. (2023). MDTD: A Multi-Domain Trojan Detector for Deep Neural Networks. *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*.

Shen, X., Chen, Z., Backes, M., Shen, Y., & Zhang, Y. (2024). "Do Anything Now": Characterising and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models. *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*.

Shen, Y., Han, Y., Zhang, Z., Chen, M., Yu, T., Backes, M., Zhang, Y., & Stringhini, G. (2022). MNEMON: Inferring Private Graph Structure from Node Embeddings. *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*.

Tramer, F., Shokri, R., San Joaquin, A., Le, H., Jagielski, M., Hong, S., & Carlini, N. (2022). Truth Serum: Poisoning Machine Learning Models to Reveal Their Secrets. *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*.

Wang, Y., Liu, Z., Luo, B., Hui, R., & Li, F. (2024). The Invisible Polyjuice Potion: An Effective Physical Adversarial Attack Against Face Recognition. *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*.

Wei, C., Zhao, M., Zhang, Z., Chen, M., Meng, W., Liu, B., Fan, Y., & Chen, W. (2023). DPMLBench: Holistic Evaluation of Differentially Private Machine Learning. *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*.

Wu, S., Bhaskar, R., Ha, A. Y. J., Shan, S., Zheng, H., & Zhao, B. Y. (2025). On the Feasibility of Poisoning Text-to-Image AI Models via Adversarial Mislabeling. *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*.

Xu, B., Dai, X., Tang, D., & Zhang, K. (2025). One Surrogate to Fool Them All: Universal, Transferable, and Targeted Adversarial Attacks with CLIP. *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*.

Yang, Y., Hu, M., Xie, X., Cao, Y., Zhang, P., Huang, Y., & Chen, M. (2025). FilterFL: Knowledge Filtering-based Data-Free Backdoor Defence for Federated Learning. *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*.

Zeng, Y., Pan, M., Just, H. A., Lyu, L., Qiu, M., & Jia, R. (2023). Narcissus: A Practical Clean-Label Backdoor Attack with Limited Information. *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*.

Yearly BreakDown

1.) 2022

Adversarial machine learning (AML) research is increasingly addressing the gap between theoretically sound attacks and practical deployment challenges. These include economic incentives, physical limitations, and enterprise architecture constraints. Historically, research emphasised attack complexity—such as sophisticated gradient-based methods—while neglecting implementation challenges faced by practitioners. Recent work published at ACM CCS 2022 demonstrates efforts to close this "significant, practical gap" (Hammoudeh & Lowd, 2022). A key challenge is adapting standard attack methods to deployed systems. Standard adversarial example generation techniques succeed against only 6.53% of real-world DNN models in Android apps due to constraints like proprietary formats and encryption (Deng et al., 2022). Deng et al. (2022) overcame these barriers through model extraction and interface reasoning, achieving a 47.35% success rate. Economic factors also shape adversary strategies: stealing a Self-Supervised Learning (SSL) encoder from Encoder-as-a-Service (EaaS) costs far less than training one from scratch (Cong et al., 2022). This vulnerability drives countermeasures like SSLGuard watermarking to protect intellectual property in commercial settings (Cong et al., 2022).

Recent research also incorporates physical and human constraints previously ignored by abstract mathematical models. Duan et al. (2022) addressed the disconnect between conventional distance metrics (*Lp* norms) and human perception by introducing a Perception-Aware Attack (PAA) framework validated

against commercial systems like YouTube's copyright detector. Muller et al. (2022) developed AttrackZone, leveraging 3D point cloud data to identify physically realisable "attack zones" for projector-based adversarial manipulation against object trackers. Vulnerability assessment is also shifting toward realistic operational models. Shen et al. (2022) proposed MNEMON, a model-agnostic graph recovery attack that assumes only access to node embeddings—a practical scenario in enterprise environments like vertical federated learning, where model interaction is prohibited. Attacks also expose vulnerabilities in efficiency-driven architectures. Li et al. (2022) found that Multi-Exit Networks (MENs) leak membership information through their exit depth configuration. Even protective measures are vulnerable: poisoning training data can dramatically reduce privacy attack costs by eliminating expensive steps like training shadow models (Tramer et al., 2022). These efforts demonstrate AML's maturation toward practical security challenges in deployed machine learning systems.

References

Source	Citation
Cong, T., He, X., & Zhang, Y. (2022).	Cong et al., 2022
Deng, Z., Chen, K., Meng, G., Zhang, X., Xu, K., & Cheng, Y. (2022).	Deng et al., 2022
Duan, R., Qu, Z., Zhao, S., Ding, L., Liu, Y., & Lu, Z. (2022).	Duan et al., 2022
Hammoudeh, Z., & Lowd, D. (2022).	Hammoudeh & Lowd, 2022
Kolluri, A., Baluta, T., Hooi, B., & Saxena, P. (2022).	Kolluri et al., 2022
Li, Z., Liu, Y., He, X., Yu, N., Backes, M., & Zhang, Y. (2022).	Li et al., 2022
Liu, Y., Zhao, Z., Backes, M., & Zhang, Y. (2022).	Liu et al., 2022
Muller, R., Man, Y., Celik, Z. B., Li, M., & Gerdes, R. (2022).	Muller et al., 2022
Shen, Y., Han, Y., Zhang, Z., Chen, M., Yu, T., Backes, M., Zhang, Y., & Stringhini, G. (2022).	Shen et al., 2022
Tramèr, F., Shokri, R., San Joaquin, A., Le, H., Jagielski, M., Hong, S., & Carlini, N. (2022).	Tramer et al., 2022
Chowdhury, A. R., Guo, C., Jha, S., & van der Maaten, L. (2022).	Chowdhury et al., 2022

2.) 2023

Following the observed divergence between theoretical adversarial machine learning (AdvML) research, which often explores complex, worst-case attacks, and the realities of production environments, recent work from ACM CCS 2023 highlights key discrepancies regarding attacker knowledge, resource constraints, and defensive limitations. Several studies reveal that traditional AdvML relies on assumptions often deemed unrealistic for practical deployment. For instance, classic clean-label backdoor attacks typically require adversaries to possess full knowledge of the victim model's training data, a condition rarely met in real-world scenarios such as crowdsourcing data collection. Similarly, existing attacks targeting ML-based Android malware detection (AMD) often rely on detailed knowledge of the feature space, model parameters, or training set, prompting the development of solutions like AdvDroidZero, which aim to succeed under a more pragmatic zero-knowledge setting. Conversely, the success of attacks requiring minimal information confirms that high-value intellectual property (IP) is vulnerable even to resource-constrained adversaries: for example, the cost of stealing large language model (LLM) decoding algorithms and proprietary hyperparameters via black-box API access is estimated to be very low, yet the information acquired is commercially valuable.

The deployment of robust AdvML solutions faces challenges stemming both from inherent computational barriers and immediate circumvention by strong adversaries. Fundamentally, complex security requirements can hinder application; for instance, verifying the robustness of general decision tree ensembles against AdvML threats is NP-hard, which motivated the creation of "verifiable learning" techniques that constrain model complexity to allow for efficient, polynomial-time verification. Moreover, implementing advanced defensive architectures does not guarantee security. Stateful Defense Models (SDMs), designed specifically to thwart query-based black-box attacks by tracking similar queries, are highly vulnerable to new adaptive strategies that exploit the information inadvertently leaked by the defense mechanism itself. Furthermore, efforts to integrate robustness often create new vulnerabilities: AdvML models enforced with security properties, designed to provide verifiable security guarantees, are found to be significantly more susceptible to model stealing attacks than traditionally trained models. This highlights a pervasive

tension between enhancing security and maintaining adequate privacy levels in deployed systems.

A significant trend addresses these practical gaps directly, focusing development toward deployable, realistic solutions. To protect high-value commercial assets, such as LLM-based code generation (LLCG) APIs, new watermarking (WM) techniques like ToSyn must be engineered to be resilient against strong, fully knowledgeable adversaries while minimizing disruption to the API's primary function and overall throughput. Similarly, addressing public safety concerns associated with emerging technologies demands focused defense strategies; research details the widespread misuse of text-to-image models to generate unsafe images and hateful memes, underscoring the urgent need for targeted safeguard tools that mitigate this scalable, real-world threat. Finally, focusing on accessibility and low computational cost is critical for wider adoption. Proposed multi-domain Trojan detection methods aim to be computationally inexpensive and broadly applicable across data domains, contrasting with past approaches that were often cost-prohibitive due to requiring specialized resources for model retraining or trigger reconstruction.

References

Apruzzese, G., Anderson, H. S., Dambra, S., Freeman, D., Pierazzi, F., & Roundy, K. A. (2022). "Real Attackers Don’t Compute Gradients": Bridging the Gap Between Adversarial ML Research and Practice. *CoRR*.

Belavadi, V., Zhou, Y., Kantarcioglu, M., & Thuraisingham, B. (2023). Attack Some while Protecting Others: Selective Attack Strategies for Attacking and Protecting Multiple Concepts. *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS ’23)*.

Calzavara, S., Cazzaro, L., Pibiri, G. E., & Prezza, N. (2023). Verifiable Learning for Robust Tree Ensembles. *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS ’23)*.

Jiang, Z., Zhang, J., & Gong, N. Z. (2023). Evading Watermark-based Detection of AI-Generated Content. *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS ’23)*.

Li, P., Xia, Y., Zhang, X., & Ji, S. (2023). Efficient Query-Based Attack against ML-Based Android Malware Detection under Zero Knowledge Setting. *Proceedings of*

the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23).

Li, Z., Wang, C., Wang, S., & Gao, C. (2023). Protecting Intellectual Property of Large Language Model-Based Code Generation APIs via Watermarks. *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*.

Miao, C., Feng, J., You, W., Shi, W., Huang, J., & Liang, B. (2023). A Good Fishman Knows All the Angles: A Critical Evaluation of Google's Phishing Page Classifier. *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*.

Naseh, A., Krishna, K., Iyyer, M., & Houmansadr, A. (2023). Stealing the Decoding Algorithms of Large Language Models. *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*.

Qu, Y., Shen, X., He, X., Backes, M., Zannettou, S., & Zhang, Y. (2023). Unsafe Diffusion: On the Generation of Unsafe Images and Hateful Memes From Text-To-Image Models. *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*.

Qin, Y., Fu, Z., Deng, C., Liao, X., Zhang, J., & Duan, H. (2023). Stolen Risks of Models with Security Properties. *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*.

Rajabi, A., Asokraj, S., Jiang, F., Niu, L., Ramasubramanian, B., Ritcey, J., & Poovendran, R. (2023). MDTD: A Multi-Domain Trojan Detector for Deep Neural Networks. *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*.

Shi, C., Zhang, T., Xu, Z., Li, S., Gao, D., Li, C., Petropulu, A., Wu, C. M., & Chen, Y. (2023). Privacy Leakage via Speech-induced Vibrations on Room Objects through Remote Sensing based on Phased-MIMO. *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*.

Wei, C., Zhao, M., Zhang, Z., Chen, M., Meng, W., Liu, B., Fan, Y., & Chen, W. (2023). DPMLBench: Holistic Evaluation of Differentially Private Machine Learning. *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*.

Wei, C., Zhao, M., Zhang, Z., Chen, M., Meng, W., Liu, B., Fan, Y., & Chen, W. (2023). DPMLBench: Holistic Evaluation of Differentially Private Machine Learning. *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*.

Yu, Z., Zhai, S., & Zhang, N. (2023). AntiFake: Using Adversarial Audio to Prevent Unauthorised Speech Synthesis. *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*.

Zeng, Y., Pan, M., Just, H. A., Lyu, L., Qiu, M., & Jia, R. (2023). Narcissus: A Practical Clean-Label Backdoor Attack with Limited Information. *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*.

3.) 2024

The current landscape in adversarial machine learning (AML) research exhibits a concerted effort to move past theoretical gradient-based and white-box assumptions toward methods reflective of real-world operational environments and resource limitations, aiming directly at the gap between academia and industry. A major theme in recent research involves demonstrating effective attacks with minimal adversarial knowledge, challenging the reliance of deployed systems on assumptions about data confidentiality or abundant computational resources. For instance, protecting proprietary assets is undermined by attacks like Neural Dehydration (Dehydra) (Lu et al., 2024), which effectively removes embedded black-box DNN watermarks even in *data-limited* settings (less than 2% of training data) or under *data-free* assumptions, proving that current intellectual property (IP) defenses are vulnerable when organizations cannot guarantee strict data isolation,,,. Similarly, in the realm of Large Language Model (LLM) applications deployed as services, sophisticated threats bypass manual effort: PLeak (Hui et al., 2024) automatically optimises *zero-query* adversarial prompts to steal confidential system prompts (IP), offering a scalable attack vector against commercial LLM applications that far surpasses manually crafted methods. Furthermore, research tracking real-world threats via JailbreakHub (Shen et al., 2024) confirms that LLM safety safeguards (like RLHF) frequently fail against continuously evolving, *in-the-wild* prompt injection and obfuscation strategies

employed by adversaries, confirming that standard defences lag behind practical attacks.

This drive toward realism extends significantly into attacks and defences concerning physical and safety-critical systems, forcing research to account for operational constraints and human factors often ignored in traditional models. In Autonomous Driving (AD), a major revelation exposing the gap involves human perception: Avara (Ma et al., 2024) employed unique VR and eye-tracking systems to demonstrate that widely adopted digital measures of attack inconspicuousness (such as L*p* norms) fail to reliably predict whether human drivers actually perceive physical adversarial attacks (PAEs) on signs, suggesting that current research metrics do not adequately address real-world safety risks for human-in-the-loop systems,,. Conversely, defenses are also adapting to leverage physical reality: VisionGuard (Han et al., 2024) successfully detects a wide range of PAEs against AD visual perception modules by exploiting their inherent *spatiotemporal inconsistency* across video frames, mitigating attacks by utilizing reliable internal kinetic data (GPS, IMU) rather than relying on complex and expensive digital certification techniques or visual features susceptible to adversarial crafting,,,,. This realism is mirrored in biometric security, where the Agile attack (Wang et al., 2024) introduces a highly stealthy, invisible infrared laser-based attack against face recognition systems, effectively bypassing visibility checks employed by typical countermeasures.

Finally, the security of modern service-based architectures (MLaaS and Distributed Learning) presents a critical gap where architectural decisions intended to improve performance or protect privacy inadvertently enable new attacks. For commercial MLaaS providers facing IP theft via model replication, research introduces defences like Beowulf (Gong et al., 2024), which actively reshapes decision regions using synthetic dummy classes and noise to make the resulting model difficult to reproduce via extraction attacks leveraging API queries. Moreover, even distributed privacy mechanisms are vulnerable: BiSR (Chen et al., 2024) reveals that incorporating Large Language Models into split learning (SL) frameworks creates significant data leakage risks, exploiting the subtle "Not-too-far" property of fine-tuning to perform successful data reconstruction despite the privacy claims of the SL architecture,,. Collectively, this research demonstrates that vulnerabilities arising from real-world constraints—be they limited data, physical limitations, human factors, or architectural complexities—are consistently

being uncovered, urging industry practice to adopt adaptive, engineering-focused solutions to bridge the identified gap.

References

(Chen et al., 2024) Guanzhong Chen, Zhenghan Qin, Mingxin Yang, Yajie Zhou, Tao Fan, Tianyu Du, and Zenglin Xu. 2024. Unveiling the Vulnerability of Private Fine-Tuning in Split-Based Frameworks for Large Language Models: A Bidirectionally Enhanced Attack. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24)*. ACM, New York, NY, USA, 15 pages. (Fang et al., 2024) Zheng Fang, Tao Wang, Lingchen Zhao, Shenyi Zhang, Bowen Li, Yunjie Ge, Qi Li, Chao Shen, and Qian Wang. 2024. Zero-Query Adversarial Attack on Black-box Automatic Speech Recognition Systems. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24)*. ACM, New York, NY, USA, 15 pages. (Gong et al., 2024) Xueluan Gong, Rubin Wei, Ziyao Wang, Yuchen Sun, Jiawen Peng, Yanjiao Chen, and Qian Wang. 2024. Beowulf: Mitigating Model Extraction Attacks Via Reshaping Decision Regions. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24)*. ACM, New York, NY, USA, 15 pages. (Han et al., 2024) Xingshuo Han, Haozhao Wang, Kangqiao Zhao, Gelei Deng, Yuan Xu, Hangcheng Liu, Han Qiu, and Tianwei Zhang. 2024. VisionGuard: Secure and Robust Visual Perception of Autonomous Vehicles in Practice. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24)*. ACM, New York, NY, USA, 18 pages. (Hui et al., 2024) Bo Hui, Haolin Yuan, Neil Gong, Philippe Burlina, and Yinzhi Cao. 2024. PLeak: Prompt Leaking Attacks against Large Language Model Applications. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24)*. ACM, New York, NY, USA, 15 pages. (Lu et al., 2024) Yifan Lu, Wenxuan Li, Mi Zhang, Xudong Pan, and Min Yang. 2024. Neural Dehydration: Effective Erasure of Black-box Watermarks from DNNs with Limited Data. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24)*. ACM, New York, NY, USA, 20 pages. (Ma et al., 2024) Xinyao Ma, Chaoqi Zhang, Huadi Zhu, L. Jean Camp, Ming Li, and Xiaojing Liao. 2024. Avara: A Uniform Evaluation System for Perceptibility Analysis Against Adversarial Object Evasion Attacks. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24)*.

ACM, New York, NY, USA, 15 pages. (Shen et al., 2024) Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2024. “Do Anything Now”: Characterising and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS ’24)*. ACM, New York, NY, USA, 15 pages. (Wang et al., 2024) Ye Wang, Zeyan Liu, Bo Luo, Rongqing Hui, and Fengjun Li. 2024. The Invisible Polyjuice Potion: An Effective Physical Adversarial Attack Against Face Recognition. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS ’24)*. ACM, New York, NY, USA, 15 pages.

4.) 2025

The research presented at ACM CCS 2025 reveals a complex but encouraging trend: adversarial machine learning (AML) research is increasingly shifting focus from purely theoretical, gradient-based attacks toward methods that exploit system-level vulnerabilities, reflect realistic adversarial constraints, and target commercial production environments. This movement explicitly seeks to bridge the gap identified in previous years by demonstrating practical harm. Several studies focus on exploiting vulnerabilities that arise when artificial intelligence (AI) systems are integrated into physical or proprietary infrastructure. For instance, **ControlLoc** (Ma et al., 2025) and research on **Asymmetry Vulnerability** (Lou et al., 2025) expose weaknesses in autonomous driving (AD) visual perception and online mapping pipelines, demonstrating success via low-cost physical-world interference and exploiting fundamental model biases in critical safety contexts. Similarly, the paper on attacking **Magika** (Nasr et al., 2025) provides a stark example of exploiting a single, publicly known ML model component (a file-type classifier) to compromise a complex, deployed production system (Gmail's malware detection pipeline), illustrating how a non-robust ML component can become the weakest link in a system-level defence. Furthermore, attackers are actively focusing on vulnerable points within industrial training pipelines, such as manipulating Vision-Language Models (VLMs) to inject **Adversarial Mislabeled Poison** (Wu et al., 2025) into training data for Text-to-Image models, and leveraging Knowledge Distillation methods to amplify malicious biases from teacher to student language models, as demonstrated by **BIASED-ROOTS** (Chaudhari et al., 2025).

The current wave of AML research simultaneously addresses practical constraints that often render classic attacks infeasible. For adversaries, the **UnivIntruder** framework (Xu et al., 2025) showcases an effective method for creating universal, transferable, and targeted adversarial examples against black-box systems, achieving high attack rates on commercial services like Google Search and GPT-4 by circumventing limitations related to high query counts or access to target training data. Meanwhile, defences are also evolving to meet the non-traditional challenges of modern AI deployment. **FilterFL** (Yang et al., 2025) and **Sylva** (Qi et al., 2025) address the failures of generalised defences in decentralised environments by developing personalised adversarial robustness solutions tailored for heterogeneous data distributions and constrained resources found in federated learning and edge computing. The advent of complex architectures, such as SuperNets, introduces new stealth opportunities, which the **VillainNet** attack (Anonymous, 2025) exploits by precisely controlling backdoor activation based on a subnetwork's operational characteristics. Finally, to combat the continuous evolution of adversarial software, **Pack-ALM** (Li et al., 2025) proposes replacing subjective, heuristic-based malware detection methods (like relying on entropy thresholds) with robust, learning-based models designed to accurately identify sophisticated and previously unseen packed binaries. Collectively, these works signal a critical convergence, where AML research is increasingly focused on generating sophisticated, practical threats against deployment scenarios while simultaneously building specialised, resource-aware defences to secure these increasingly complex systems.

References

- Anonymous. 2025. VillainNet: Targeted Poisoning Attacks Against SuperNets Along the Accuracy-Latency Pareto Frontier. In Proceedings of (CCS '25). ACM, New York, NY, USA.
- Chaudhari, Harsh, Jamie Hayes, Matthew Jagielski, Ilia Shumailov, Milad Nasr, and Alina Oprea. 2025. Cascading Adversarial Bias from Injection to Distillation in Language Models. In Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS '25), Taipei, Taiwan. ACM, New York, NY, USA.

- Cohen, Stav, Ron Bitton, and Ben Nassi. 2025. Here Comes The AI Worm: Unleashing Zero-click Worms that Target GenAI-Powered Applications. In Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS '25), Taipei, Taiwan. ACM, New York, NY, USA.
- Li, Shijia, Jiang Ming, Lanqing Liu, Longwei Yang, Ni Zhang, and Chunfu Jia. 2025. Adversarially Robust Assembly Language Model for Packed Executables Detection. In Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS '25), Taipei. ACM, New York, NY, USA.
- Lou, Yang, Haibo Hu, Qun Song, Qian Xu, Yi Zhu, Rui Tan, Wei-Bin Lee, and Jianping Wang. 2025. Asymmetry Vulnerability and Physical Attacks on Online Map Construction for Autonomous Driving. In Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS '25), Taipei, Taiwan. ACM, New York, NY, USA.
- Ma, Chen, Ningfei Wang, Zhengyu Zhao, Qian Wang, Qi Alfred Chen, and Chao Shen. 2025. ControlLoc: Physical-World Hijacking Attack on Visual Perception in Autonomous Driving. In Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS '25), Taipei, Taiwan. ACM, New York, NY, USA.
- Nasr, Milad, Yanick Fratantonio, Luca Invernizzi, Ange Albertini, Loua Farah, Alex Petit-Bianco, Andreas Terzis, Kurt Thomas, Elie Bursztein, and Nicholas Carlini. 2025. Evaluating the Robustness of a Production Malware Detection System to Transferable Adversarial Attacks. In Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS '25), Taipei, Taiwan. ACM, New York, NY, USA.
- Qi, Tianyu, Lei Xue, Yufeng Zhan, and Xiaobo Ma. 2025. Sylva: Tailoring Personalized Adversarial Defense in Pre-trained Models via Collaborative Fine-tuning. In Proceedings of 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS '25). ACM, New York, NY, USA.
- Wu, Stanley, Ronik Bhaskar, Anna Yoo Jeong Ha, Shawn Shan, Haitao Zheng, and Ben Y. Zhao. 2025. On the Feasibility of Poisoning Text-to-Image AI Models via Adversarial Mislabeling. In Proceedings of the 2025 ACM SIGSAC

Conference on Computer and Communications Security (CCS '25), Taipei, Taiwan. ACM, New York, NY, USA.

- Xu, Binyan, Xilin Dai, Di Tang, and Kehuan Zhang. 2025. One Surrogate to Fool Them All: Universal, Transferable, and Targeted Adversarial Attacks with CLIP. In Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS '25), Taiwan. ACM, New York, NY, USA.
- Xu, Xiaoyun, Zhuoran Liu, Stefanos Koffas, and Stjepan Picek. 2025. Towards Backdoor Stealthiness in Model Parameter Space. In Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS '25), Taipei, Taiwan. ACM, New York, NY, USA.
- Yang, Yanxin, Ming Hu, Xiaofei Xie, Yue Cao, Pengyu Zhang, Yihao Huang, and Mingsong Chen. 2025. FilterFL: Knowledge Filtering-based Data-Free Backdoor Defence for Federated Learning. In Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS '25), Taipei, Taiwan. ACM, New York, NY, USA.