# Bridging the Gap Between Theory and Practice in Adversarial Machine Learning: A Systematic Cross-Venue Analysis of 454 Papers (2022–2025)

Madhav Khanal
Rollins College
mkhanal@rollins.edu

JJ Jasser
Rollins College
jjasser@rollins.edu

## Abstract

Machine learning systems are increasingly employed in security-critical contexts. However, difference still exists between academic work in adversarial machine learning studies and the requirements of practical applications in the wild. This systematic review examines 454 articles in four leading security conferences, namely ACM CCS, IEEE S&P, NDSS, and USENIX Security, between 2022 and 2025. Using a conceptual structure developed in "Real Attackers Don't Compute Gradients" by Apruzzese et al., we assess the extent to which current studies have bridged the gap between theory and practice.

We found that most papers don't test on deployment systems in 94.7% of studies, 67.8% require rare gradient access in real-world settings, 80.4% query budgets are beyond real-world constraints, and 63.2% require white-box models. We identified five regions that require more focus for real-world deployment: model specifications, model access to gradients, validation on real systems, diversity regions beyond image classification, and economics and human-centric focus. Our study, from 2022 to 2025, verifies the gap within the identified regions. Our goal is to demonstrate how real-world attacks deviate from research in general while also acknowledging the significant contributions that each paper has made. We conclude with directions for researchers, conferences, and funding bodies to build upon the substantial contribution of prior research and foster research that bridges theoretical rigour with real-world security needs.

**Keywords:** adversarial machine learning, theory-practice gap, security research, systematic review, threat modeling

# Contents

# 1    Introduction

Machine learning has become an essential component of modern software systems. It is used in facial recognition at border crossings [1], fraud detection in financial services [2], navigation for autonomous vehicles [3], and content moderation on large social platforms [4]. As a result, ML models increasingly influence decisions that affect millions of people each day. This widespread deployment has intensified scrutiny of adversarial vulnerabilities, defined here as inputs or interactions crafted to cause models to behave in unintended and potentially harmful ways [5].

Adversarial machine learning (AML) encompasses the study of attacks against ML systems and defenses to improve their security and robustness [6]. Over the past decade, researchers have demonstrated a range of attacks across the ML pipeline: evasion attacks that craft malicious inputs to cause misclassification at inference time [7, 8], poisoning attacks that inject corrupted training data to compromise model behavior [9, 10], and privacy attacks that extract sensitive information through model queries or inspections [11, 12]. In response, a substantial body of work has proposed defenses intended to improve robustness, privacy, and resistance to adversarial manipulation [13, 14].

**Why This Gap Matters: Real-World Stakes** The consequences of adversarial vulnerabilities extend beyond academic benchmarks. In November 2025, Anthropic disclosed what it assessed with high confidence to be the first large-scale state-sponsored cyber espionage campaign executed predominantly by an AI agent [15]. Attackers manipulated Claude Code into autonomously conducting reconnaissance, exploit generation, and data exfiltration against approximately thirty global targets, and the system performed an estimated 80–90% of the operational workload with minimal human oversight. Critically, this incident did not rely on gradient-based attacks or white-box model access; instead, it exploited deployment-level assumptions about agentic autonomy, tool access via the Model Context Protocol [16], and safeguard circumvention through task decomposition [17]. This case exemplifies a broader pattern documented in Section 3: real adversaries exploit system integration points, human workflows, and economic leverage rather than computing optimal $L_p$-bounded perturbations [18, 19].

The economic and societal stakes are substantial. Adversarial attacks impose direct financial costs: deepfake-enabled fraud has caused documented losses reaching millions of dollars in individual incidents [20], while data extraction attacks can expose proprietary training data worth substantial investment [12]. Organizations face barriers to deploying adversarial defenses due to unclear ownership of ML security risks, insufficient tooling for adversarial testing, and limited understanding of realistic threat models [21, 22]. Industry surveys reveal widespread uncertainty about how to assess adversarial risks in production systems [23]. Meanwhile, regulatory frameworks increasingly mandate robustness guarantees for high-stakes ML deployments [24], creating compliance costs and potential penalties, yet practitioners report that academic defenses often prove impractical under operational constraints [25].

**The Apruzzese Critique** In 2022, Apruzzese and colleagues published "Real Attackers Don't Compute Gradients" [26], a critical analysis highlighting a persistent gap between academic AML research and practical security concerns. Their core observation was that much of the literature assumes attackers possess capabilities rarely available in real-world settings: white-box access (complete knowledge of model architectures, parameters, and training procedures [27]), the ability to compute gradients through target models [28], effectively unlimited query budgets [29], and an absence of monitoring or rate limiting [30]. Through case studies of actual ML attacks, they demonstrated that real adversaries more often rely on simpler strategies, including basic input manipulation, exploitation of system-level weak-

nesses, or social engineering, when these are sufficient to achieve their objectives. This critique raises a fundamental question about whether academic research priorities align with operational security needs.

**Research Questions** Based on the Apruzzese critique, we conduct this systematic review to answer the following questions:

- **RQ1:** To what extent has adversarial ML research addressed practical deployment constraints since the 2022 Apruzzese critique?
- **RQ2:** What patterns distinguish research that bridges the theory–practice gap from work that remains laboratory-focused?
- **RQ3:** How do emerging threats, including large language models and foundation models, affect the alignment between academic research and operational security needs?

**Positioning in Prior Work** This review builds on a growing body of meta-analyses examining the practical relevance of security research. Kumar et al. [22] conducted interviews with 28 organizations and found widespread uncertainty about deploying adversarial defenses. Grosse et al. [23] surveyed 271 industry practitioners, revealing systematic overestimation of realistic attacker capabilities in academic work. Arp et al. [25] performed a meta-analysis of security research, documenting reliance on simplified laboratory setups that inadequately account for adaptive adversaries. Apruzzese et al. [26] synthesized these concerns through case studies demonstrating that real ML attacks rarely resemble academic threat models.

Our work extends this line of inquiry in four ways. First, we provide the first systematic cross-venue analysis of the post-Apruzzese period (2022–2025), examining whether the critique catalyzed methodological changes. Second, we introduce a quantitative Gap Score framework applied uniformly across 454 papers, enabling statistical comparison of research practices across venues and time. Third, we document the emergence of new threat landscapes, particularly agentic LLM misuse and multimodal attacks, that introduce adversarial challenges distinct from traditional computer vision settings. Fourth, we provide actionable recommendations tailored to four stakeholder groups rather than general observations.

**Scope and Boundaries** To address these questions, we analyze 454 adversarial ML papers published between 2022 and 2025 at four leading security venues: ACM CCS (118 papers), IEEE S&P (79 papers), NDSS (49 papers), and USENIX Security (208 papers). We selected these venues because they represent top-tier security research outlets that explicitly emphasize real-world impact and practical relevance in their calls for papers and review criteria. The 2022–2025 timeframe captures the research community's response in the post-critique period, with sufficient temporal coverage to identify trends while maintaining methodological tractability.

Our analysis excludes papers from general ML conferences (NeurIPS, ICML, ICLR), workshop-only publications, and preprints. This exclusion is deliberate. Security-focused venues face different review criteria and author incentives than ML theory venues, where advancing state-of-the-art benchmark performance may supersede deployment considerations [31, 32]. Security conference review criteria explicitly emphasize threat model realism, evaluation on operational systems, and consideration of attacker incentives [33]. By focusing on security conferences, we examine research that ostensibly prioritizes practical security impact, making the theory–practice gap more salient when present.

Each paper is evaluated along dimensions adapted from Apruzzese et al. [26] and extended to capture deployment realism: threat model assumptions (white-box, gray-box, or black-box adversary knowl-

edge [27]), reliance on gradient access (whether attacks require backpropagation through the target model), query budget requirements (number of model inferences needed), computational cost (GPU requirements and training time), validation setting (evaluation on real production systems versus static benchmarks), and consideration of economic or organizational constraints (costs, incentives, and operational barriers). These dimensions are operationalized into a Gap Score (0–6), where 0 indicates full alignment with practical constraints and 6 indicates reliance on all six idealized assumptions (detailed in Section 4).

**Preview of Key Findings** Our analysis reveals that the theory–practice gap identified in 2022 remains substantial. Key findings include:

- *Limited real-world evaluation:* Only 5.3% of papers (24 of 454) evaluate attacks or defenses on deployed systems (operational production environments with real users, monitoring, and operational constraints), while the remainder rely on offline datasets and simulated laboratory environments.
- *Persistent gradient dependence:* 67.8% of papers require gradient information from target models (access to model internals enabling backpropagation-based attacks [34]), with minimal decline from 2022 (68.2%) to early 2025 (65.9%).
- *White-box dominance:* 63.2% assume white-box adversaries with complete model knowledge, despite industry surveys indicating such access is rare in practice [23]. In contrast, black-box settings (query-only access without model internals [? ]) better reflect most deployment scenarios.
- *High query budgets:* Among papers involving model queries, 80.4% assume budgets exceeding 1000 queries, often ignoring cost constraints (commercial API pricing [? ]), rate limiting [30], and anomaly detection systems [35].
- *Average Gap Score of 3.17:* Most papers incorporate approximately half of the idealized assumptions we track, with only 10.5% scoring 0–1 (closely aligned with deployment constraints).

Despite these gaps, we observe encouraging trends. Code release rates have increased, improving reproducibility [36]. Emerging work explores query-efficient and gradient-free methods [37, 38]. There is growing attention to LLM-specific threats including jailbreaking [39] and prompt injection [40], along with modest diversification beyond computer vision domains into text [41], audio [42], and malware detection [43].

**Contributions** This review makes the following contributions:

- *First post-Apruzzese cross-venue systematic analysis:* We provide the first comprehensive assessment of how 454 papers published across four top-tier security venues (ACM CCS, IEEE S&P, NDSS, USENIX Security, 2022–2025) address the theory–practice gap following the influential Apruzzese critique [26].
- *Quantitative Gap Score framework:* We introduce and apply a systematic coding framework with a six-dimensional Gap Score, enabling statistical comparison of practical relevance across venues, years, and research focus (attack versus defense), extending prior qualitative meta-analyses [22, 25].
- *Documentation of real-world adversarial incidents:* We synthesize documented cases of ML attacks in deployed systems (Section 3), including the first large-scale agentic AI cyber campaign [15], prompt injection vulnerabilities [40], and deepfake fraud [20], illustrating the sociotechnical nature of adversarial threats.

- *Emergence analysis of the LLM threat landscape:* We characterize how foundation models (large-scale pre-trained models adapted to diverse tasks [44]) and agentic systems (autonomous AI systems with tool access and iterative decision-making [45]) introduce adversarial challenges—including jailbreaking [39], prompt injection [40], and multimodal attacks [46]—that diverge from traditional threat models (Section 7).
- *Actionable stakeholder-specific recommendations:* We provide targeted guidance for researchers (evaluation methodology), conference organizers (review criteria), practitioners (deployment considerations), and funders (bridging infrastructure), grounded in our empirical findings.

**Who This Review Serves** This work is intended for multiple stakeholder groups. Researchers designing new attacks or defenses can use our findings to align evaluation methodologies with deployment realities and identify underexplored but practically relevant research directions. Conference program committees and reviewers can reference our quantitative analysis when assessing the practical relevance of submissions and designing review criteria that reward realistic threat modeling. Industry practitioners can use our gap analysis to critically evaluate whether proposed academic defenses address their operational threat models and constraints, avoiding adoption of methods validated only under idealized assumptions. Funding agencies can use our recommendations to structure programs encouraging academia–industry collaboration, realistic benchmarks, and bridging infrastructure. Collectively, we hope this review catalyzes methodological reflection and incremental alignment between adversarial ML research and operational security needs.

## 2 Background: Foundations of Adversarial Machine Learning

### 2.1 The Discovery of Adversarial Vulnerabilities

The modern study of adversarial machine learning began with the important work of Szegedy et al. [47]. They showed that small, barely noticeable changes to input data could lead state-of-the-art deep learning models to produce wrong outputs with high confidence. These changes were not just random mistakes; they were systematic vulnerabilities that could transfer between different models. Adversarial examples created for one model often worked against other models trained separately. This ability to transfer highlighted the nature of the vulnerability and sparked further research.

Goodfellow et al. [48] suggested that the linear behavior of neural networks, instead of their nonlinearity or overfitting, was the main cause of their vulnerability to adversarial attacks. They introduced the Fast Gradient Sign Method (FGSM), which made it easier to generate adversarial inputs using a single gradient step. The speed and effectiveness of FGSM set a standard for later research, as gradient-based attacks became the common approach.

These foundational studies shaped the direction of adversarial ML research in three significant ways. First, they assumed that adversaries had white-box access to models, including details about their architecture and gradient information. Second, they focused on imperceptible changes measured by specific distance metrics. Third, they highlighted optimization-based methods as the best way to generate adversarial inputs. While these frameworks allowed for thorough analysis, they also shifted research away from the practical realities of real-world attacks.

## 2.2 The Arms Race: Attacks and Defenses

After these initial findings, researchers came up with more advanced gradient-based attacks. Moosavi-Dezfooli et al. [49] introduced DeepFool, which calculated minimal changes iteratively to cross decision boundaries with smaller, more accurate modifications than FGSM. Carlini and Wagner [34] approached the generation of adversarial examples as a complex optimization problem tailored to various distance metrics ($L_0$, $L_2$, and $L_\infty$), achieving high success rates against certain defenses, including defensive distillation [50].

This led to an arms race between attack and defense research. Defenses were proposed and tested against known attacks, only to be later bypassed by adaptive gradient-based methods. Carlini and Wagner [51] systematically defeated ten detection-based defenses, showing that evaluations limited to known attacks fell short when adversaries adapted their strategies to account for defense mechanisms.

Adversarial training became a significant defense approach. Madry et al. [13] redefined adversarial robustness through a min-max optimization framework, training models to perform well against the worst-case changes within set limits. Tramer et al. [52] introduced ensemble adversarial training to enhance black-box robustness. Research on certified robustness [53] aimed to provide formal guarantees that models would remain accurate within certain perturbation limits, although these methods often involved trade-offs between accuracy and robustness and required considerable computing power.

## 2.3 Expanding Threat Landscape

The field then expanded beyond evasion attacks to include privacy violations and threats during training. Membership inference attacks [11, 54] determine whether specific data points were part of model training. Carlini et al. [55] argued for assessing such attacks at low false positive rates, which reflect the serious nature of privacy issues.

Training-time attacks interfere with the learning process itself. Backdoor attacks introduce hidden triggers that cause specific misbehaviors when activated, with recent studies showing attacks on self-supervised learning [56], along with distribution-preserving methods that avoid detection [57]. Poisoning attacks that corrupt training data demonstrate that even small amounts of malicious samples can significantly harm model performance [58].

Physical-world attacks took gradient-based techniques beyond digital settings. Kurakin et al. [59] showed that adversarial examples remained effective when printed and photographed. Later work [60] developed robust physical adversarial examples against traffic sign recognition systems in autonomous vehicles, raising questions about whether real-world attackers would use such complex methods when simpler options might work.

## 2.4 The Theory-Practice Gap

As the theoretical level of adversarial ML research increased, worries grew about whether this work addressed actual security needs in the real world. Kumar et al. [22] interviewed practitioners from 28 organizations, discovering that the industry lacked practical tools for dealing with adversarial ML threats. Grosse et al. [61] surveyed 271 industrial practitioners, revealing that while academic threat models were theoretically valid, research often overestimated what attackers could do, especially regarding access to training data and query limits.

Mink et al. [21] identified organizational barriers to deploying defenses through interviews with ML practitioners. These barriers included a lack of motivation from institutions, difficulty assessing AML risk, and competing business priorities. Practitioners often saw security as outside their expertise, revealing mismatches between job definitions in ML and security duties.

Apruzzese et al. [26] crystallized these issues through real-world case studies showing that actual ML system breaches often involved simple tactics rather than complex gradient-based attacks: input manipulation without gradients, exploiting system weaknesses instead of model-specific faults, operational gaps, and social engineering. This historical trend encourages a systematic analysis of recent publications to see if current research is addressing the theory-practice gap or continuing past patterns.

## 2.5 Types of Adversarial Attacks

Adversarial attacks generally fall into three categories based on their goals:

**Evasion attacks** manipulate inputs during inference to cause misclassification. The attacker changes an input—such as an image, malware sample, or network packet—so that the model gives an incorrect prediction. These attacks focus on the inference phase and represent the most studied threat type.

**Poisoning attacks** corrupt the training process. By adding malicious examples to training data, attackers can lead models to learn wrong behaviors or establish backdoors—hidden triggers that cause specific failures when activated. A compromised model might classify most inputs correctly but struggle with ones containing a particular pattern.

**Privacy attacks** extract sensitive information. Membership inference attacks find out if specific persons were included in the training data. Model extraction attacks aim to replicate model functions through repeated queries. Data reconstruction attacks try to retrieve actual training examples.

## 2.6 Threat Models and Their Implications

A threat model outlines what an adversary can do and what they know. This aspect is a major area where academic research diverges from real-world usage.

**White-box access** suggests the attacker knows everything about the model: its architecture, parameters (weights), and possibly its training data. With white-box access, attackers can compute gradients—the mathematical changes needed to alter inputs and impact model outputs. Many academic attacks assume white-box access because it simplifies the process of optimizing attacks.

**Black-box access** means the attacker can only query the model and observe outputs. This situation is closer to real-world uses where models operate as web services or are embedded in applications. The attacker cannot see internal model details; they can only submit inputs and get predictions.

**Gray-box access** represents middle-ground scenarios. The attacker might know the model architecture but not the specific parameters or might have access to a similar training dataset.

Apruzzese et al. stress that real-world attackers rarely have white-box access [26]. Production models usually have security measures in place. Yet, as our analysis shows, 63.2% of papers in our dataset assume white-box access, which may not reflect what actually happens in deployment.

# 3 Methodology

## 3.1 Data Collection

We systematically reviewed all papers with adversarial ML focus published at four top-tier security venues from 2022 through 2025. These venues were selected because they represent primary publication outlets for security-focused ML research and have historically shaped the field's direction.
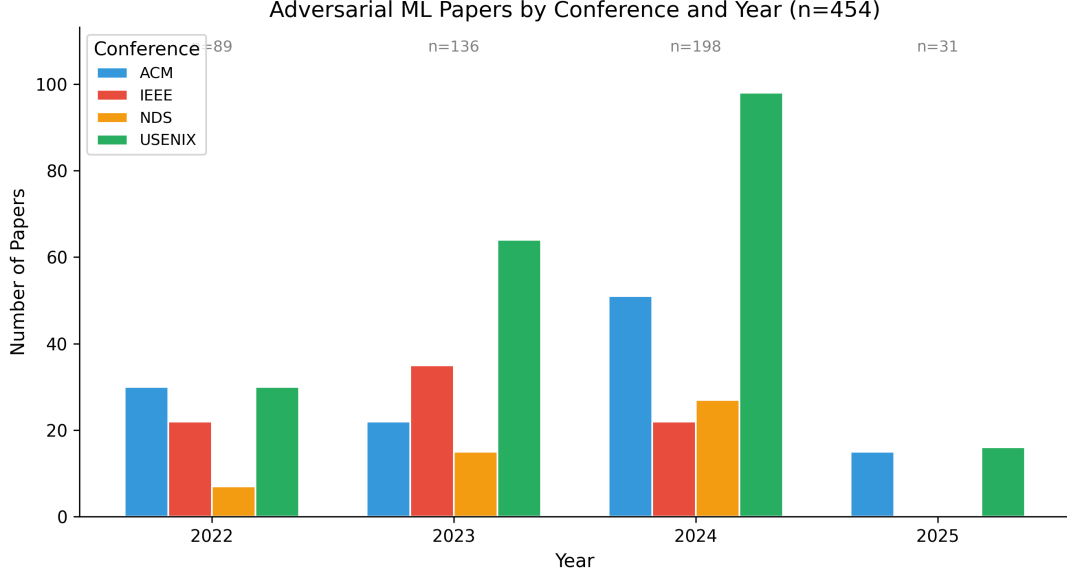


Figure 1: Dataset overview showing the distribution of 454 papers across four security conferences (ACM CCS, IEEE S&P, NDSS, USENIX Security) from 2022 to 2025.

Our dataset comprises 454 papers: ACM CCS contributed 118 papers (26.0%), IEEE S&P contributed 79 papers (17.4%), NDSS contributed 49 papers (10.8%), and USENIX Security contributed 208 papers (45.8%). The temporal distribution includes 89 papers from 2022, 136 from 2023, 198 from 2024, and 31 from 2025 (partial year at time of analysis).

## 3.2 Coding Framework

Each paper was evaluated across multiple dimensions designed to capture practical relevance:

- **Research Focus (G1):** Whether the paper primarily proposes attacks (60%), defenses (39%), or both (2%).
- **Attack Type (G2):** Classification as evasion (48%), poisoning (20%), privacy (23%), or multiple types (9%).
- **Data Domain (G4):** The input modality: images (65%), text (11%), audio (7%), malware (6%), or other (12%).
- **Threat Model (T1):** Assumed adversary access: white-box (63.2%), black-box (34.1%), or gray-box (2.6%).
- **Gradient Requirements (Q1):** Whether the approach requires gradient access.
- **Query Budget (Q2):** High (>1000 queries), low, or none.
- **Real System Testing (G7):** Whether validation occurred on deployed systems.

- **Code Release (G6):** Whether code was released publicly.

## 3.3 Gap Score Framework

To quantify the theory-practice gap, we developed a 6-point "Gap Score" summing binary indicators of assumptions that may limit practical applicability:

1. Requires white-box access (vs. black/gray-box)
2. Requires gradient computation
3. Assumes high query budget ($>1000$ queries)
4. Requires substantial computation (GPU-level resources)
5. No testing on deployed systems
6. No consideration of economic factors

Higher scores indicate greater distance from practical deployment considerations. A paper scoring 0 would employ realistic threat models, require minimal resources, validate on production systems, and consider economic constraints. A paper scoring 6 would represent a primarily theoretical contribution with limited immediate applicability to deployed systems.

# 4 Quantitative Findings
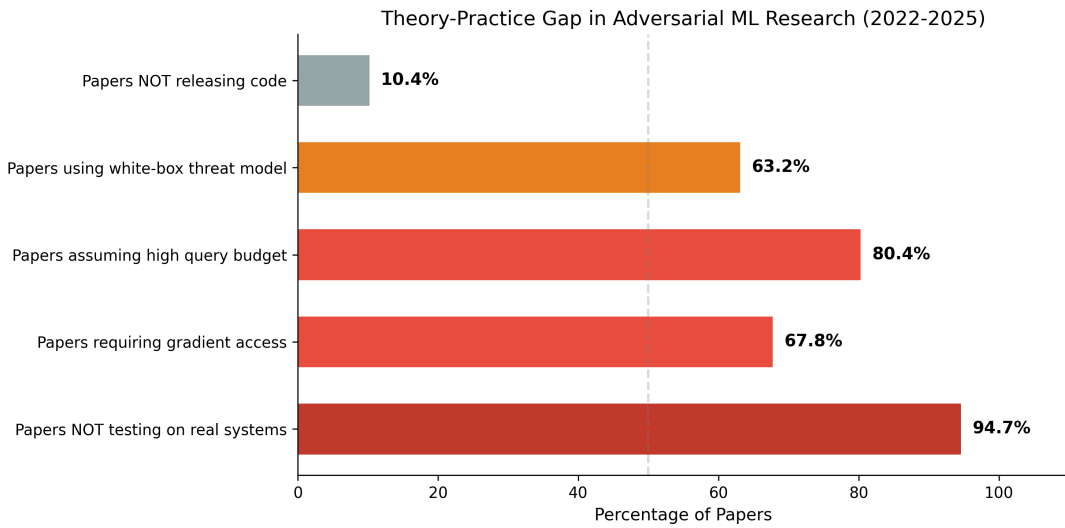


Figure 2: Overview of the theory-practice gap showing the percentage of papers exhibiting each characteristic: no real-world testing (94.7%), high query budget (80.4%), gradient dependency (67.8%), white-box access (63.2%), and no code release (10.4%).

Our analysis identifies several areas where research practices diverge from deployment realities. Figure 2 summarizes the key indicators across all 454 papers.

## 4.1 Real-World Validation

The most pronounced finding concerns real-world validation: **only 5.3% of papers (24 of 454) evaluate on actual deployed systems**. The remaining 94.7% validate exclusively on research benchmarks,

simulated environments, or research prototypes.
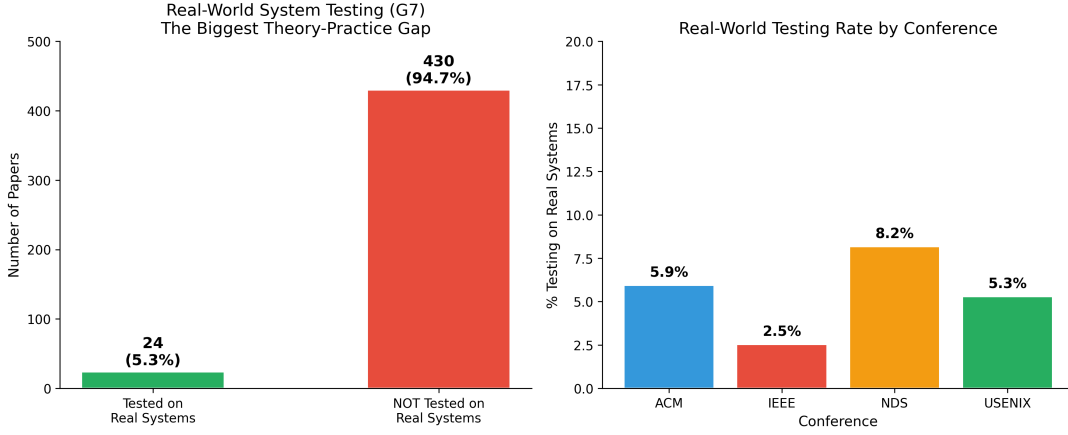


Figure 3: Real-world testing rate: 5.3% of papers validate on deployed systems.

This pattern has implications because deployment introduces constraints absent from controlled experiments. Production systems employ proprietary model formats and encryption. They integrate ML components into complex security pipelines where multiple components interact. They face hardware constraints, latency requirements, and regulatory compliance obligations. Research validated only in laboratory settings may require substantial adaptation for deployment.

The few papers that do validate on real systems often reveal differences between laboratory and field performance. Nayan et al. [62] conducted a systematic review of on-device ML model extraction attacks, finding that many proposed academic attacks proved difficult to reproduce, performed less effectively on production models, or introduced unacceptable computational and energy costs. Similarly, Layton et al. [63] demonstrated that deepfake detection research employs metrics and dataset distributions that may lead to overestimation of detector efficacy.

Duan et al. [64] validated perception-aware attacks against YouTube's copyright detection system, representing one of the rare examples of testing against commercial infrastructure. Their work demonstrated that attacks optimized in simulation required adaptation to succeed in practice.

## 4.2 Gradient Dependency

Apruzzese et al.'s critique centered on the observation that "real attackers don't compute gradients." Our analysis indicates this pattern persists: **67.8% of papers require gradient access**, despite this capability being unavailable against most production systems.
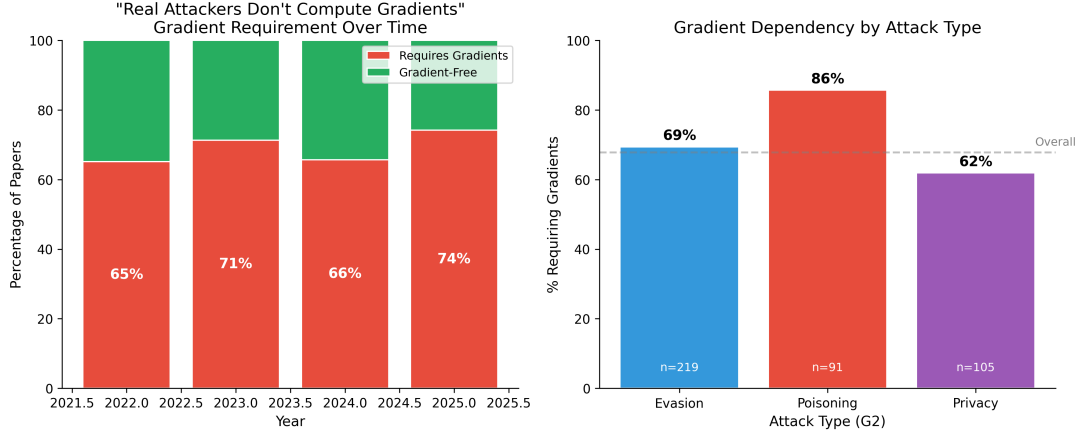
Figure 4: Gradient dependency analysis showing limited change over time (68.2% in 2022 to 65.9% in 2025) and variation by attack type.

Temporal analysis reveals limited change in this dimension. Gradient dependency has remained relatively stable: 68.2% in 2022, 69.1% in 2023, 66.8% in 2024, and 65.9% in 2025. While individual papers have explored gradient-free approaches, the overall distribution has not shifted substantially.

Some research has pioneered gradient-free methods in specific domains. The Universal Robustness Evaluation Toolkit (URET) from Eykholt et al. [38] formulates adversarial generation as a graph exploration problem, seeking sequences of domain-specific, functionality-preserving transformations rather than relying on differentiable feature spaces. This framework enables evaluation of systems processing inputs like malware binaries or tabular data where semantic and functional correctness must be maintained during perturbation.

Similarly, practical LLM jailbreak attacks demonstrate that effective attacks need not be gradient-based. Liu et al. [65] and Yu et al. [66] showed that jailbreaking large language models can be effective even when executed via strategically crafted natural language prompts, illustrating the potential of accessible black-box attacks compared to complex gradient optimization.

## 4.3 Threat Model Assumptions

White-box access remains the predominant assumption: **63.2% of papers assume adversaries have complete knowledge of model architecture and parameters**. Only 34.1% consider black-box scenarios more closely matching deployment conditions, and 2.6% examine gray-box settings.
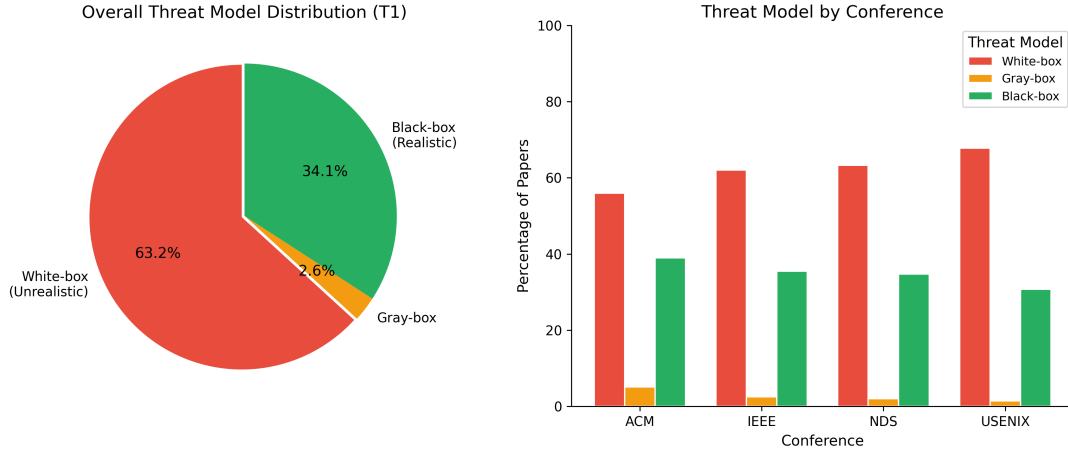
Figure 5: Distribution of threat model assumptions across all papers, showing prevalence of white-box assumptions.

This pattern contrasts with industrial practice. As Grosse et al. [61] document, academic studies often operate under assumptions of attacker access—such as extensive access to internal models, parameters, or training data—that do not reflect the security controls present in production environments.

The foundational meta-analysis by Arp et al. [67] identified methodological patterns in security research, including reliance on laboratory-only evaluation and deployment of threat models that may not account for adaptive adversaries. Our analysis suggests these patterns persist in the 2022–2025 literature.

## 4.4 Query Budget Assumptions

Even papers employing black-box threat models often assume substantial query access. **80.4% of papers assume high query budgets (>1000 queries)**, which may be impractical when commercial APIs implement rate limiting, repeated queries trigger anomaly detection, and real-time constraints limit iterative optimization.
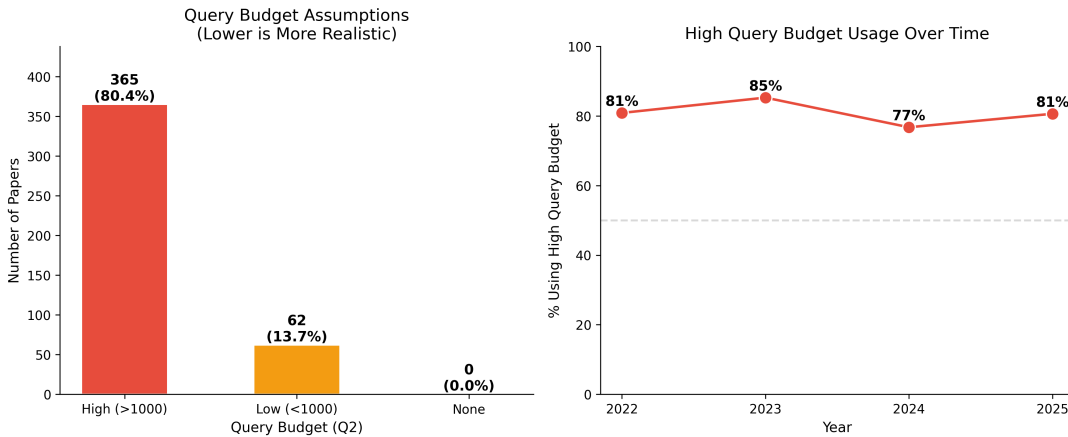


Figure 6: Query budget assumptions showing 80.4% of papers assume high query budgets.

Some recent work has addressed query efficiency. HARDBEAT from Tao et al. [68] generates triggers requiring knowledge only of the final predicted label (hard-label) and minimal queries, address-

ing restrictions imposed by commercial services. BounceAttack from Wan et al. [69] demonstrates query-efficient decision-based attacks. However, these remain exceptions rather than the predominant approach.

## 4.5 Domain Distribution

Adversarial ML research exhibits notable domain concentration: **65% of papers focus exclusively on image data**. Text receives 11% of attention, audio 7%, malware 6%, with all other domains combined representing 12%.
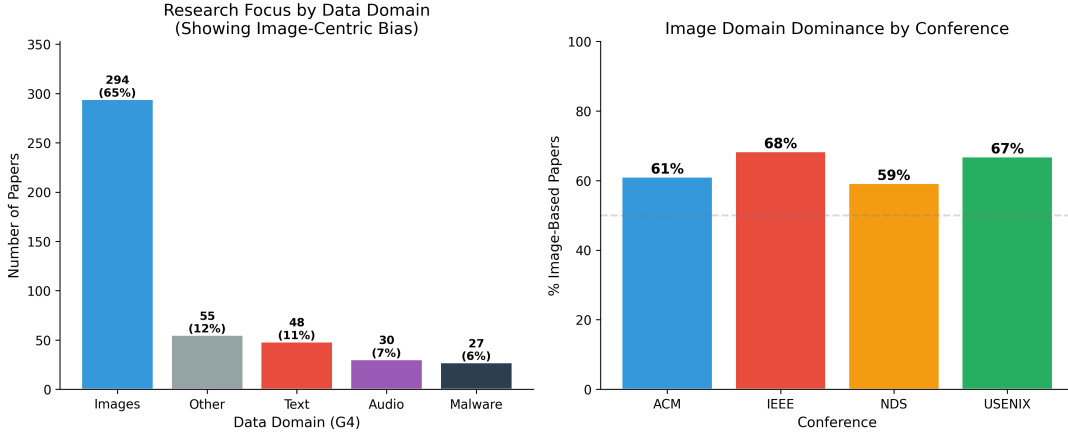


Figure 7: Data domain distribution revealing concentration on image classification in adversarial ML research.

This concentration creates potential blind spots. Financial systems process tabular data where adversarial perturbations cannot be measured by pixel distances. As Kireev et al. [70] observe for fraud detection, the meaningful constraint is not visual imperceptibility but rather the quantifiable financial cost or utility an adversary must expend. $L_p$ norms may be less informative for tabular financial data; what matters is whether fraudulent transactions remain economically viable.

## 4.6 Code Availability

One dimension shows encouraging results: **89.6% of papers release their code**. This represents substantial commitment to reproducibility and exceeds code release rates in many other fields.

However, this introduces a nuanced consideration: code designed for research datasets may not transfer directly to production environments without substantial re-engineering. High code release rates support reproducibility while not necessarily indicating deployment readiness.

## 4.7 Gap Score Distribution
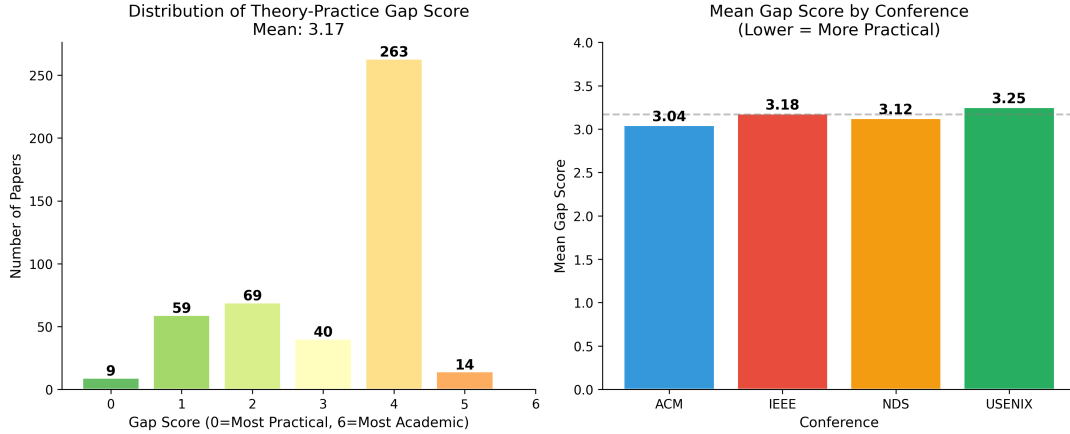


Figure 8: Distribution of Gap Scores showing the typical paper (mean 3.17/6) incorporates roughly half of the measured assumptions.

The mean Gap Score across all 454 papers is **3.17 out of 6**, indicating the typical paper incorporates approximately half of the assumptions we measured. The distribution peaks at scores of 3 to 4, with approximately 10.5% of papers achieving scores of 0 to 1 that would indicate closer alignment with deployment considerations.

Conference-level analysis reveals similar patterns across venues: ACM CCS averages 3.04, NDSS 3.12, IEEE S&P 3.18, and USENIX 3.25. No venue has established itself as substantially more practice-focused than others; the patterns we observe appear consistent across the security research community.

# 5 Thematic Findings

Beyond quantitative metrics, thematic analysis across venues reveals structural patterns in how adversarial ML research is conducted and evaluated.

## 5.1 The Utility-Robustness Trade-off

A significant consideration for deployment is the relationship between security guarantees and system performance. Research from USENIX Security (2022–2025) particularly illuminates this tension.

Xiang et al. [71] found that defensive proposals achieving certifiable robustness against adversarial patches frequently yielded reduced clean classification accuracy, which may discourage real-world deployment. By 2024, Xiang et al. [72] documented that certifiably robust defenses in computer vision require 10 to 100 times more inference-time computation than undefended models, presenting computational challenges for practical deployment.

This pattern extends beyond vision systems. Ahmed et al. [73] found that defenses against malicious activations in voice assistants typically affect natural accuracy. Similarly, widely used privacy-preserving techniques like DP-SGD may compromise model utility to achieve privacy guarantees [74, 75].

Some recent work demonstrates progress in addressing this trade-off. PatchCleanser introduced a double-masking approach compatible with any image classifier, achieving high certified robustness

while preserving state-of-the-art clean accuracy [71]. MIST offers a pathway toward robust security by strategically limiting overfitting only to the most membership-vulnerable training instances [76]. CAMP training demonstrated that provable adversarial robustness in deep reinforcement learning need not sacrifice certified expected return, which may be valuable for safety-critical robotics applications [77].

## 5.2 Evaluation Methodology

Research across venues reveals considerations regarding how attacks and defenses are evaluated.

### 5.2.1 Distance Metrics

Standard evaluation measures adversarial perturbation size using $L_p$ norms—mathematical measures of distance between original and perturbed inputs. However, as Carlini et al. [55] document, conventional distance metrics like $L_p$ norms, which measure pixel-level differences, do not reliably predict whether humans perceive adversarial perturbations.

This consideration was empirically examined by the Avara framework [78], which used VR environments and eye-tracking to study whether drivers notice adversarial traffic signs. They found that $L_p$ norms do not reliably predict whether human drivers notice adversarial perturbations: attacks deemed "imperceptible" by mathematical standards may be immediately obvious to a driver, while attacks violating $L_p$ constraints might go unnoticed in realistic driving conditions.

### 5.2.2 Privacy Evaluation

Privacy attacks present particular evaluation challenges. Carlini et al. [55] observe that privacy is fundamentally a worst-case concern: a defense succeeds only if it protects all individuals, not just the majority. Yet membership inference attacks are typically evaluated using average-case metrics (overall accuracy, AUC) that may mask severe privacy leakage for specific individuals.

## 5.3 Physical Deployment Considerations

A consideration for conventional AML research concerns the gap between digital adversarial examples and the complex physical conditions governing real-world perception systems.

### 5.3.1 Physical Attack Challenges

Digital perturbations optimized in simulation may perform differently when deployed physically due to environmental factors: distance and viewing angle variations, illumination changes, sensor noise, and compression artifacts. NDSS research has particularly emphasized this consideration.

Jia et al. [60] developed robust physical adversarial example pipelines tested against production autonomous vehicles running YOLO v5 traffic sign recognition. Their work required accounting for real-road conditions that simulation may not capture.

Physical attack research at USENIX has expanded beyond vision systems. Liu et al. [79] designed physically realizable 3D adversarial objects capable of deceiving X-ray prohibited item detection, requiring optimization for shape rather than color or texture and accounting for complex object overlap in

luggage. Cao et al. [80] demonstrated Physical Removal Attacks using focused laser spoofing to selectively remove LiDAR point cloud data on autonomous vehicles. The "Tubes Among Us" research [81] demonstrated analog adversarial attacks where adversaries manipulate voice signals using simple tubes to bypass speaker recognition, effectively bypassing established digital artifact detection methods.

By 2025, physical attack research has embraced increasingly practical scenarios. "Shadow Hack" [82] exploits LiDAR weaknesses using ordinary non-reflective materials placed on roads, requiring no specialized equipment. ATKSCOPES [83] demonstrates rapid evasion against real-world perceptual hashing algorithms by dynamically adapting to victim systems.

### 5.3.2 System Integration

Research increasingly recognizes that integrating ML models into complex systems introduces considerations that isolated model analysis may miss. Debenedetti et al. [84] reveal that system-level components such as training data filters or output monitoring may introduce privacy side channels exploitable by adaptive adversaries, potentially affecting provable differential privacy guarantees.

Nasr et al. [85] demonstrated this through examining how exploiting Google's Magika file-type classifier affects Gmail's malware detection pipeline—a single non-robust component can affect an entire security pipeline.

Chen et al. [43] document that security mechanisms integrated solely at the framework level may not persist once models are compiled into optimized executables for deployment. Defenses may need to be embedded within the DL compiler pipeline to persist through deployment, a requirement often absent from academic work.

## 5.4 Attack Evolution

Despite methodological considerations about assumptions, research has progressively explored more practical attack vectors.
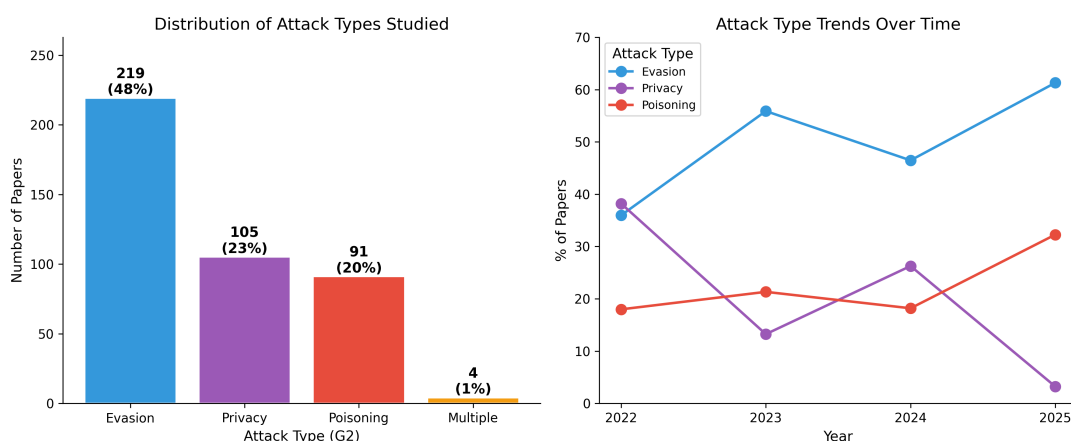


Figure 9: Attack type distribution showing evasion attacks (48%) predominate, with trends over time.

Early practical attacks (2022) demonstrated physically realizable approaches like the "frustum attack," which leverages environmental context to compromise automotive sensor fusion in black-box settings [86].

By 2023–2024, attacks increasingly emphasized simplicity and accessibility. The effectiveness of jailbreaking LLMs through natural language prompts, even by users without ML expertise, illustrated the potential of accessible black-box attacks compared to complex gradient optimization.

The 2025 landscape shows attacks prioritizing stealth and persistence. MergeBackdoor [87] reveals supply chain considerations where seemingly benign upstream models pass security checks but activate malicious backdoors upon merging with other components. Guo et al. [88] describe persistent backdoor strategies targeting stable neuronal components, ensuring exploits survive continuous parameter updates in continual learning systems.

## 5.5    Defense Evolution

Defensive research has evolved from generic solutions toward specialized, interpretable, and context-aware mechanisms.

Blacklight [35] detected and mitigated black-box query-based attacks against MLaaS by leveraging the observation that iterative optimization produces highly similar queries, providing defense against persistent attackers that bypass account-based security measures.

Recent defenses demonstrate increased sophistication. JBShield [89] moves beyond heuristics for LLM protection by using the Linear Representation Hypothesis to identify and manipulate "toxic" and "jailbreak" concepts within hidden states. SafeSpeech [90] proactively affects voice data during training to make synthesized audio less usable, achieving robustness against voice cloning. DeBackdoor [91] addresses deployment constraints by providing backdoor detection effective under black-box access, data scarcity, and pre-deployment inspection limitations.

## 5.6    Economic and Organizational Factors

Academic research has given limited attention to the economic and organizational dimensions of adversarial ML.

### 5.6.1    Economic Considerations

ACM CCS research has highlighted how commercial ML services create incentives for IP theft that may affect protective mechanisms. Cong et al. [92] document that extracting pre-trained encoders may cost substantially less than training from scratch. Lu et al. [93] show that Neural Dehydration can remove watermarks using less than 2% of training data. The economic asymmetry—where model extraction attacks may cost orders of magnitude less than defenses or the assets they protect—remains an area requiring additional attention.

### 5.6.2    Organizational Considerations

Beyond technical considerations, qualitative research reveals organizational factors affecting industry adoption. Mink et al. [21] found that ML practitioners often lack institutional motivation and resources to understand and mitigate adversarial threats, frequently viewing security and machine learning as disconnected fields.

This organizational separation results in lower prioritization of adversarial evaluations before deployment and limited visibility into monitoring for active attacks. Defenses may remain unimplemented

due to isolation between ML and security teams or because competing business priorities outweigh the cost and time needed for robust implementation.

# 6 Temporal Trends

A central question motivates this review: has the research community responded to calls for greater practical relevance? Our temporal analysis provides insight into this question.
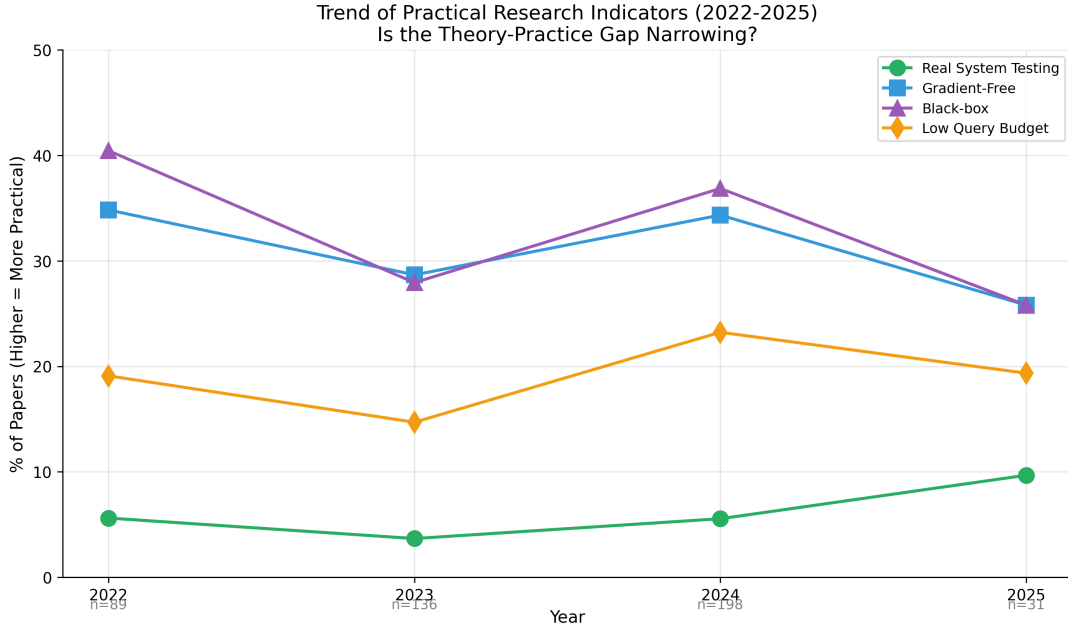


Figure 10: Temporal trends showing the theory-practice gap from 2022 to 2025, with limited change in key metrics.

## 6.1 Stable Patterns

**Real-world testing** remains at approximately 5% across all years. Despite calls for deployment validation, the research community has not substantially shifted toward testing on production systems.

**Gradient dependency** has remained relatively stable at 67–69%. There has been limited movement toward gradient-free approaches at the aggregate level.

**White-box assumptions** show no substantial reduction (63% in 2025 vs. 64% in 2022). Threat model distributions have remained consistent.

**Domain distribution** persists with image data representing 65% of papers throughout the study period.

**Economic analysis** remains below 11% in all years.

## 6.2 Areas of Modest Progress

Query efficiency shows some progress, with high-budget assumptions declining from 80.4% to approximately 76% by 2025. Some researchers have developed query-efficient attacks, though these represent a minority of approaches.

Awareness of attacker knowledge limitations has increased, with growing acknowledgment that adversaries often lack full system knowledge. However, this awareness has not yet translated into substantial shifts in threat modeling practices.

## 6.3 Emerging Research Areas

The 2024–2025 period has seen growth in LLM security research, introducing new considerations. Jailbreak attacks evolve rapidly relative to RLHF defenses [94]. Prompt injection against commercial LLM services poses practical concerns. Adversarial training, the standard defense approach, remains computationally intensive for broad deployment.

# 7 Cross-Venue Analysis
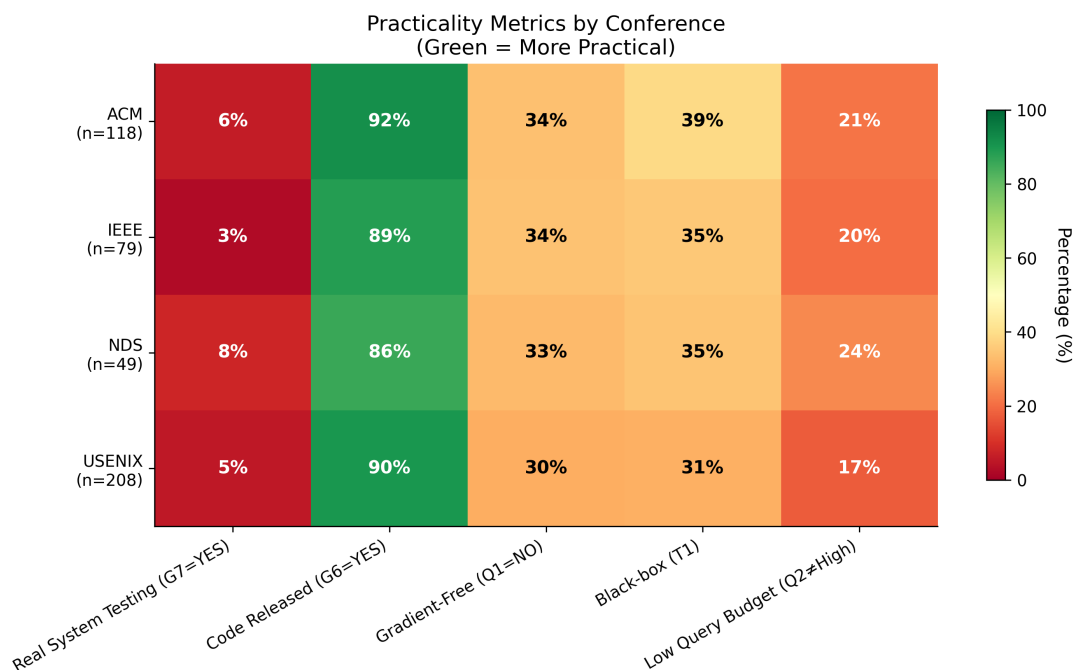


Figure 11: Conference comparison heatmap showing similar patterns across all venues.

While all venues share similar theory-practice gap patterns, each has developed distinctive emphases that collectively illuminate different facets of the research landscape.

## 7.1 ACM CCS: Economics and System Integration

ACM CCS research uniquely emphasizes business logic, usability constraints, and intellectual property protection. Contributions include demonstrating how economic incentives shape the threat landscape [92, 93], documenting relationships between mathematical metrics and human perception [78], analyzing system-level integration considerations [85], and examining architectural constraints in emerging deployment patterns.

## 7.2   IEEE S&P: Efficiency and Formal Guarantees

IEEE S&P research emphasizes computational efficiency, privacy-utility trade-offs, and certified scalability. The venue has advanced understanding of threat modeling considerations and resource assumptions [55], documented accuracy-privacy trade-offs [95], and examined scalability considerations for certified robustness deployment [96].

## 7.3   NDSS: Physical Constraints and Operational Realities

NDSS research emphasizes system-level constraints, distributed training considerations, and physical deployment realities. Contributions include documenting gaps between digital perturbations and physical deployment [60], developing metrics for non-image domains [70], analyzing distributed training heterogeneity [97], and studying adversarial dynamics where defenses create exploitable side channels.

## 7.4   USENIX Security: Real-World Validation and Domain Diversity

USENIX Security research emphasizes testing on deployed systems and domain-specific constraints. The venue has documented tensions between theoretical rigor and deployment constraints [71, 72], analyzed trade-offs involving computational cost, regulatory compliance, and usability [73], studied adversarial dynamics when defenses encounter adaptive attackers in production, and examined domain-specific requirements [38].

## 7.5   Attack versus Defense Papers
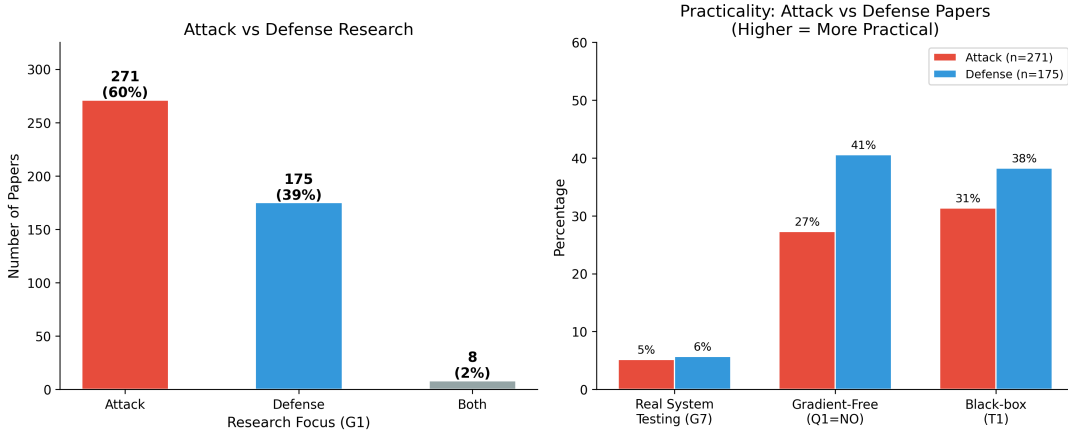


Figure 12: Comparison of practicality metrics between attack-focused and defense-focused papers.

Our analysis reveals that defense papers demonstrate somewhat higher alignment with practical considerations than attack papers on average. Defense papers show 6% real-system testing versus 5% for attacks, 41% gradient-free approaches versus 27%, and 38% black-box focus versus 31%. However, both categories remain distant from deployment-ready research.

# 8 Recommendations

Bridging the theory-practice gap requires coordinated effort across the research ecosystem. We offer recommendations for multiple stakeholder groups.

## 8.1 For Researchers

### 8.1.1 Methodological Considerations

**Validate on real systems when feasible.** Partner with industry to evaluate against production deployments. Even limited real-world testing can reveal constraints not apparent in laboratory settings.

**Employ realistic threat models.** Consider black-box access as a default assumption. When white-box assumptions are necessary, justify them explicitly and acknowledge their implications.

**Expand domain coverage.** Extend research beyond image classification to tabular data, graphs, time-series, and multimodal systems where practical deployment is occurring.

**Consider economic factors.** Analyze attack costs, defender resources, and return-on-investment trade-offs. Security decisions involve economic considerations.

**Include human evaluation.** Test perceptibility against actual humans rather than relying solely on mathematical metrics.

### 8.1.2 Evaluation Practices

Evaluate privacy using true-positive rate at low false-positive rates, not only average accuracy or AUC. For non-image domains, define domain-appropriate perturbation constraints. Report query budgets realistically based on what commercial APIs actually permit. Measure computational requirements in practical units such as wall-clock time and computational cost.

## 8.2 For Venues and Program Committees

**Encourage real-world validation.** At minimum, request authors to justify why real-system testing is infeasible if omitted.

**Consider artifact badges for deployability.** Recognize work tested on production systems with explicit acknowledgment.

**Develop evaluation guidelines.** Request papers to explicitly address threat model realism, query budget constraints, gradient requirements, domain-appropriate metrics, economic considerations, and adaptive adversary testing.

**Value negative results.** Papers demonstrating that attacks fail under realistic constraints provide valuable information that positive-result publication preferences may suppress.

## 8.3 For Industry Practitioners

**Interpret academic results in deployment context.** Adjust risk assessments for deployment realities rather than accepting laboratory success rates directly.

**Anticipate adaptation requirements.** Budget for re-engineering before research prototypes become production-ready.

**Prioritize defenses against realistic threats:** black-box attacks, low query budget scenarios, economically motivated adversaries, and domain-specific constraints.

**Consider contributing data.** Anonymized information about real attack attempts could help ground academic research in deployment reality.

## 8.4 For Funding Agencies

Support industry-academic partnerships with real-world validation components. Fund deployment studies rather than only paper publication. Incentivize replication studies that validate academic claims against production systems. Consider red team / blue team programs with realistic constraints.

# 9 Limitations and Threats to Validity

## 9.1 Scope Limitations

This review focuses on four security venues and may not capture patterns at ML conferences (NeurIPS, ICML, ICLR) where different norms may apply. Publication preferences may suppress negative results; attacks that fail under realistic constraints or defenses that prove impractical are less likely to be published. Our 2022–2025 window may not capture longer-term trends.

## 9.2 Coding Limitations

The Gap Score reduces complex trade-offs to binary decisions, potentially oversimplifying nuanced situations. Manual coding introduces subjectivity, particularly for papers spanning multiple categories. The definition of "real system testing" may vary; some papers test on emulated production environments that share some but not all deployment constraints.

## 9.3 Generalizability

Security venues may actually be more practice-focused than typical computer science venues given their traditional emphasis on real-world threats. Our findings may therefore underestimate the theory-practice gap in the broader ML research community.

# 10 Conclusion

This systematic review involving 454 research papers published from 2022–2025 at four prestigious security conferences bears evidence of some beneficial contributions by the research community of adversarial machine learning studies in spite of the theory-practice gap noticed by Apruzzese et al.

The quantitative trends are striking:

- 94.7% of the papers don't test on real systems
- 67.8% need gradients which might not be accessible in real life
- 80.4% assume query budgets that could exceed deployment limits
- 63.2% assume white-box access that could not be provided in deployment situations

Thematic analysis shows other considerations emerging. $L_p$ norms may not model human perception well. Robustness may not generalize to production-quality settings. Defense systems do not necessarily take into consideration economic losses and use ease-of-use constraints. Assessment metrics are often averages for the worst-case scenario characteristics.

Among the many contributions of the research community is the theoretical foundation and pioneering research in adversarial phenomena. However, to move forward, accomplishing the goal of closing the theory-practice gap will necessitate structural changes in conferences that reward verification of effectiveness in the real world, assumptions of reasonable threats in research, industry inputs of deployment results, and distribution of funding that emphasizes effectiveness along with innovation.

However, the efficacy of ever more widespread ML-dependent systems is contingent on the ongoing involvements of the community of ML researchers with both the theory and the practicalities. Our observations indicate that those practicalities can, through shifts within publishing, evaluation, and collaboration, help ensure that community-driven ML adversarial research and development work increasingly serves a practical, security context.

# References

[1] Patrick Grother, Mei Ngan, and Kayee Hanaoka. Face recognition vendor test (frvt) part 3: Demographic effects. Technical Report NISTIR 8280, National Institute of Standards and Technology, 2019.

[2] Andrea Dal Pozzolo, Olivier Caelen, Reid A Johnson, and Gianluca Bontempi. Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications*, 41 (10):4915–4928, 2014.

[3] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

[4] Tarleton Gillespie. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. Yale University Press, 2018.

[5] Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

[6] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael P Wellman. SoK: Security and privacy in machine learning. In *IEEE European Symposium on Security and Privacy*, pages 399–414, 2016.

[7] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[8] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[9] Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *International Conference on Machine Learning*, pages 1467–1474, 2012.

[10] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. BadNets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.

[11] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy*, pages 3–18, 2017.

[12] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *USENIX Security Symposium*, pages 2633–2650, 2021.

[13] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[14] Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320, 2019.

[15] Anthropic. Disrupting the first reported AI-orchestrated cyber espionage campaign. https://www.anthropic.com/research/disrupting-the-first-reported-ai-orchestrated-cyber-espionage-campaign, 2025. Anthropic Research Blog.

[16] Anthropic. Model context protocol. https://modelcontextprotocol.io, 2024.

[17] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? In *Advances in Neural Information Processing Systems*, 2024.

[18] Justin Gilmer, Ryan P Adams, Ian Goodfellow, David Andersen, and George E Dahl. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 3–14, 2018.

[19] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *International Conference on Learning Representations*, 2018.

[20] The Guardian. Finance worker pays out $25 million after video call with deepfake "chief financial officer". https://www.theguardian.com/, 2024. News article.

[21] Jaron Mink et al. "Security is Not My Field, I'm a Stats Guy": A qualitative root cause analysis of barriers to adversarial machine learning defenses in industry. In *Proceedings of the 32nd USENIX Security Symposium*, 2023.

[22] Ram Shankar Siva Kumar, Magnus Nyström, John Lambert, Andrew Marshall, Mario Goertzel, Andi Comissoneru, Matt Swann, and Sharon Xia. Adversarial machine learning – industry perspectives. In *IEEE Security and Privacy Workshops*, pages 69–75, 2020.

[23] Kathrin Grosse, Lukas Bieringer, Tarek R Besold, and Hongxin Hu. Towards more practical threat models in artificial intelligence security. In *Proceedings of the 33rd USENIX Security Symposium*, 2024.

[24] European Commission. Proposal for a regulation on a european approach for artificial intelligence. Technical Report COM(2021) 206 final, European Commission, 2021.

[25] Daniel Arp, Erwin Quiring, Feargus Pendlebury, Alexander Warnecke, Fabio Pierazzi, Christian Wressnegger, Lorenzo Cavallaro, and Konrad Rieck. Dos and don'ts of machine learning in computer security. In *Proceedings of the 31st USENIX Security Symposium*, 2022.

[26] Giovanni Apruzzese, Hyrum S Anderson, Savino Dambra, David Freeman, Fabio Pierazzi, and Kevin A Roundy. "real attackers don't compute gradients": Bridging the gap between adversarial ml research and practice. *arXiv preprint arXiv:2212.14315*, 2022.

[27] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. In *arXiv preprint arXiv:1902.06705*, 2019.

[28] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *ACM Asia Conference on Computer and Communications Security*, pages 506–519, 2017.

[29] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 15–26, 2017.

[30] Mika Juuti, Sebastian Szyller, Samuel Marchal, and N Asokan. PRADA: Protecting against DNN model stealing attacks. In *IEEE European Symposium on Security and Privacy*, pages 512–527, 2019.

[31] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems*, pages 2503–2511, 2015.

[32] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.

[33] Cormac Herley and Paul C Van Oorschot. SoK: Science, security and the elusive goal of security as a scientific pursuit. In *IEEE Symposium on Security and Privacy*, pages 99–120, 2014.

[34] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pages 39–57, 2017.

[35] Huiying Li, Shawn Shan, Emily Wenger, Jiayun Zhang, Haitao Zheng, and Ben Y Zhao. Blacklight: Scalable defense for neural networks against query-based black-box attacks. In *Proceedings of the 31st USENIX Security Symposium*, 2022.

[36] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *Journal of Machine Learning Research*, 22(164):1–20, 2021.

[37] Guanhong Tao et al. Hard-label black-box universal adversarial patch attack. In *Proceedings of the 32nd USENIX Security Symposium*, 2023.

[38] Kevin Eykholt et al. URET: Universal robustness evaluation toolkit (for evasion). In *Proceedings of the 32nd USENIX Security Symposium*, 2023.

[39] Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[40] Yupei Liu, Jinyuan Jia, Runpeng Geng, Jinyuan Jia, and Neil Zhenqiang Gong. Formalizing and benchmarking prompt injection attacks and defenses. In *USENIX Security Symposium*, 2024.

[41] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 2153–2162, 2019.

[42] Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *IEEE Security and Privacy Workshops*, pages 1–7, 2018.

[43] Yanzuo Chen et al. OBSAN: An out-of-bound sanitizer to harden DNN executables. In *Proceedings of the 32nd USENIX Security Symposium*, 2023.

[44] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[45] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):1–26, 2024.

[46] Xiangyu Qi, Kaixuan Zeng, Tinghao Xie, Pin-Yu Chen, Peter Henderson, Matt Fredrikson, and Siheng Zhu. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21527–21536, 2024.

[47] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[48] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[49] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. DeepFool: A simple and accurate method to fool deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2574–2582, 2016.

[50] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE Symposium on Security and Privacy*, pages 582–597, 2016.

[51] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *ACM Workshop on Artificial Intelligence and Security*, pages 3–14, 2017.

[52] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018.

[53] Linyi Li, Tao Xie, and Bo Li. SoK: Certified robustness for deep neural networks. In *IEEE Symposium on Security and Privacy*, 2023.

[54] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In *ACM Conference on Computer and Communications Security*, pages 3093–3106, 2022.

[55] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *Proceedings of the 2022 IEEE Symposium on Security and Privacy*, 2022.

[56] Jinyuan Jia, Yupei Liu, and Neil Zhenqiang Gong. BadEncoder: Backdoor attacks to pre-trained encoders in self-supervised learning. *arXiv preprint arXiv:2108.00352*, 2021.

[57] Guanhong Tao, Zhenting Wang, Shiwei Feng, Guangyu Shen, Shiqing Ma, and Xiangyu Zhang. Distribution preserving backdoor attack in self-supervised learning. In *IEEE Symposium on Security and Privacy*, 2024.

[58] Florian Tramèr, Reza Shokri, Ayrton San Joaquin, Hoang Manh Le, Matthew Jagielski, Sanghyun Hong, and Nicholas Carlini. Truth serum: Poisoning machine learning models to reveal their secrets. In *ACM Conference on Computer and Communications Security*, pages 2779–2792, 2022.

[59] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *International Conference on Learning Representations Workshop*, 2017.

[60] Yaomin Jia et al. Physical adversarial attack on vehicle detector in the carla simulation environment. In *Proceedings of the 2022 Network and Distributed System Security Symposium*, 2022.

[61] Kathrin Grosse, Lukas Bieringer, Tarek R Besold, and Hongxin Hu. Towards more practical threat models in artificial intelligence security. In *Proceedings of the 33rd USENIX Security Symposium*, 2024.

[62] Tanvir Nayan, Qian Guo, Mohammed Al Duniawi, Marcus Botacin, Selcuk Uluagac, and Ruimin Sun. SoK: All you need to know about on-device ML model extraction - the gap between research and practice. In *Proceedings of the 33rd USENIX Security Symposium*, 2024.

[63] Samuel Layton, Tyler Tucker, Dominik Olszewski, Kevin Warren, Kevin Butler, and Patrick Traynor. SoK: The good, the bad, and the unbalanced: Measuring structural limitations of deepfake media datasets. In *Proceedings of the 33rd USENIX Security Symposium*, 2024.

[64] Rui Duan et al. Perception-aware attack: Creating adversarial music via reverse engineering. In *Proceedings of the 31st USENIX Security Symposium*, 2022.

[65] Haoyu Liu, Yuxuan Zhang, Zihan Zhao, Yunjie Dong, Guozhu Meng, and Kai Chen. Making them ask and answer: Jailbreaking large language models in few queries via disguise and reconstruction. In *Proceedings of the 33rd USENIX Security Symposium*, 2024.

[66] Zhiyuan Yu, Xiaogeng Liu, Shunning Liang, Zach Cameron, Chaowei Xiao, and Ning Zhang. Don't listen to me: Understanding and exploring jailbreak prompts of large language models. In *Proceedings of the 33rd USENIX Security Symposium*, 2024.

[67] Daniel Arp, Erwin Quiring, Feargus Pendlebury, Alexander Warnecke, Fabio Pierazzi, Christian Wressnegger, Lorenzo Cavallaro, and Konrad Rieck. Dos and don'ts of machine learning in computer security. In *Proceedings of the 31st USENIX Security Symposium*, 2022.

[68] Guanhong Tao et al. Hard-label black-box universal adversarial patch attack. In *Proceedings of the 32nd USENIX Security Symposium*, 2023.

[69] Yuying Wan et al. BounceAttack: A query-efficient decision-based black-box attack. In *Proceedings of the 2024 IEEE Symposium on Security and Privacy*, 2024.

[70] Klim Kireev et al. On the robustness of machine learning models beyond adversarial settings. In *Proceedings of the 2023 Network and Distributed System Security Symposium*, 2023.

[71] Chong Xiang, Saeed Mahloujifar, and Prateek Mittal. PatchCleanser: Certifiably robust defense against adversarial patches for any image classifier. In *Proceedings of the 31st USENIX Security Symposium*, 2022.

[72] Chong Xiang, Tong Wu, Sihui Dai, Jonathan Petit, Suman Jana, and Prateek Mittal. PATCHCURE: Improving certifiable robustness, model utility, and computation efficiency of adversarial patch defenses. In *Proceedings of the 33rd USENIX Security Symposium*, 2024.

[73] Shimaa Ahmed, Ilia Shumailov, Nicolas Papernot, and Kassem Fawaz. Towards more robust keyword spotting for voice assistants. In *Proceedings of the 31st USENIX Security Symposium*, 2022.

[74] Yugeng Liu, Rui Wen, Xinlei He, Ahmed Salem, Zheng Zhang, Michael Backes, Emiliano De Cristofaro, Mario Fritz, and Yang Zhang. ML-Doctor: Holistic risk assessment of inference attacks against machine learning models. In *Proceedings of the 31st USENIX Security Symposium*, 2022.

[75] Xinyu Tang, Saeed Mahloujifar, Liwei Song, Virat Shejwalkar, Milad Nasr, Amir Houmansadr, and Prateek Mittal. Mitigating membership inference attacks by self-distillation through a novel ensemble architecture. In *Proceedings of the 31st USENIX Security Symposium*, 2022.

[76] Jiacheng Li et al. MIST: Defending against membership inference attacks through membership-invariant subspace training. In *Proceedings of the 33rd USENIX Security Symposium*, 2024.

[77] Dingwen Wang et al. CAMP in the odyssey: Provably robust reinforcement learning with certified radius maximization. In *Proceedings of the 34th USENIX Security Symposium*, 2025.

[78] Yueqi Ma et al. Avara: Measuring human perception of adversarial traffic sign examples using virtual and augmented reality. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*, 2024.

[79] Siyuan Liu et al. X-Adv: Physical adversarial object attacks against x-ray prohibited item detection. In *Proceedings of the 32nd USENIX Security Symposium*, 2023.

[80] Yulong Cao et al. You can't see me: Physical removal attacks on lidar-based autonomous vehicles driving frameworks. In *Proceedings of the 32nd USENIX Security Symposium*, 2023.

[81] Shimaa Ahmed et al. Tubes among us: Analog attack on automatic speaker identification. In *Proceedings of the 32nd USENIX Security Symposium*, 2023.

[82] Ryunosuke Kobayashi et al. Invisible but detected: Physical adversarial shadow attack and defense on lidar object detection. In *Proceedings of the 34th USENIX Security Symposium*, 2025.

[83] Yizheng Zhang et al. ATKSCOPES: Multiresolution adversarial perturbation as a unified attack on perceptual hashing and beyond. In *Proceedings of the 34th USENIX Security Symposium*, 2025.

[84] Edoardo Debenedetti et al. Privacy side channels in machine learning systems. In *Proceedings of the 33rd USENIX Security Symposium*, 2024.

[85] Milad Nasr et al. Exploiting component vulnerabilities in production ML pipelines. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, 2025.

[86] R Spencer Hallyburton, Yupei Liu, Yulong Cao, Z Morley Mao, and Miroslav Pajic. Security analysis of camera-lidar fusion against black-box attacks on autonomous vehicles. In *Proceedings of the 31st USENIX Security Symposium*, 2022.

[87] Lu Wang et al. From purity to peril: Backdooring merged models from "harmless" benign components. In *Proceedings of the 34th USENIX Security Symposium*, 2025.

[88] Zhen Guo et al. Persistent backdoor attacks in continual learning. In *Proceedings of the 34th USENIX Security Symposium*, 2025.

[89] Shenyi Zhang et al. JBShield: Defending large language models from jailbreak attacks through activated concept analysis and manipulation. In *Proceedings of the 34th USENIX Security Symposium*, 2025.

[90] Zhiyuan Zhang et al. SafeSpeech: Robust and universal voice protection against malicious speech synthesis. In *Proceedings of the 34th USENIX Security Symposium*, 2025.

[91] Dusan Popovic et al. DeBackdoor: A deductive framework for detecting backdoor attacks on deep models with limited data. In *Proceedings of the 34th USENIX Security Symposium*, 2025.

[92] Tianshuo Cong et al. SSLGuard: A watermarking scheme for self-supervised learning pre-trained encoders. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2022.

[93] Yifan Lu et al. 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*, 2024.

[94] Xinyue Shen et al. Dynamic attention-based approaches for llm robustness. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*, 2024.

[95] Shahbaz Rezaei et al. On the accuracy-privacy trade-off of deep ensembles. In *Proceedings of the 2023 IEEE Symposium on Security and Privacy*, 2023.

[96] Linyi Li et al. SoK: Certified robustness for deep neural networks. In *Proceedings of the 2023 IEEE Symposium on Security and Privacy*, 2023.

[97] Phillip Rieger et al. DeepSight: Mitigating backdoor attacks in federated learning through deep model inspection. In *Proceedings of the 2022 Network and Distributed System Security Symposium*, 2022.

# A    Complete Paper Analysis Dataset

The complete analysis of all 454 papers is available in the supplementary CSV file: `all_conferences_analysis_`

The dataset includes the following columns for each paper:

- Year, Conference, Filename, Title, Authors
- G1 (Focus), G2 (Attack Type), G3 (ML Type), G4 (Data Domain)
- G5 (Economics), G6 (Code Release), G7 (Real System Testing)
- T1 (Threat Model), T2 (Training Data Access)
- Q1 (Gradient Requirements), Q2 (Query Budget), Q3 (Computation)
- Gap indicator flags and Traditional Score

Benchmark meanings (aligned with the CSV fields used in the table):

- G1 Focus: atk (attack), def (defense), both.
- G2 Attack Type: Evasion, Poisoning, Privacy, Multiple.
- G3 ML Type: DL, Traditional, Both.
- G4 Data Domain: Images, Text, Audio, Malware, Other.
- G5 Economics mentioned: YES/NO.
- G6 Code released: YES/NO.
- G7 Real system testing: YES/NO.
- T1 Threat model: White-box, Gray-box, Black-box.
- T2 Training data access: Full, Partial, None.
- Q1 Requires gradients: YES/NO.
- Q2 Query budget: High ($>1000$), Low ($<1000$), None.
- Q3 Computation: High (GPU), Low (CPU).
- Traditional_Score (Gap Score 0–6): sum of six impractical-assumption flags (higher = less practical).

For full per-paper details (all 24 columns), please see the CSV file. Below is a compact, bordered summary table focused on the most critical taxonomy fields. Binary fields are rendered as checkmarks (yes) or blanks (no); threat/access levels are abbreviated (W/B/G/P, F/P, H/L).

# Table 1: Compact summary of papers (key taxonomy fields).

| Year | Venue | Paper (1st author) | G1 | G2 | G3 | G4 | G5 | G6 | G7 | T1 | T2 | Grad | Qry | Comp | WB | Trad. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2022 | ACM | Using | def | Poisoning | DL | Images | ✓ | ✓ | × | G | F | ✓ | H | L | 0 | 3 |
| 2022 | ACM | Reverse-Engineering | atk | Evasion | DL | Audio | ✓ | ✓ | × | B | | × | H | L | 0 | 2 |
| 2022 | ACM | Tianshuo | def | Privacy | DL | Images | ✓ | ✓ | × | B | | × | L | L | 0 | 1 |
| 2022 | ACM | CISPA | atk | Privacy | DL | Other | ✓ | × | × | B | | × | L | L | 0 | 2 |
| 2022 | ACM | CISPA | atk | Privacy | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2022 | ACM | School | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2022 | ACM | Physical | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2022 | ACM | Framework | def | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2022 | ACM | Florian | atk | Poisoning | DL | Images | ✓ | × | × | W | F | ✓ | H | L | 1 | 5 |
| 2022 | ACM | via | atk | Privacy | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2022 | ACM | ExamplesforLeader | def | Evasion | DL | Images | ✓ | ✓ | × | W | P | ✓ | H | L | 1 | 4 |
| 2022 | ACM | Harnessing | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2022 | ACM | Feature | atk | Privacy | DL | Other | ✓ | × | ✓ | B | | × | L | L | 0 | 1 |
| 2022 | ACM | Imperial | atk | Privacy | Both | Other | ✓ | ✓ | × | B | | × | H | L | 0 | 2 |
| 2022 | ACM | Learning | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2022 | ACM | CISPA | atk | Evasion | DL | Text | ✓ | ✓ | × | B | | × | H | L | 0 | 2 |
| 2022 | ACM | Amrita | def | Poisoning | DL | Images | ✓ | × | × | B | | ✓ | H | L | 0 | 4 |
| 2022 | ACM | A | def | Evasion | Both | Text | ✓ | ✓ | × | B | | × | L | L | 0 | 1 |
| 2022 | ACM | CISPA | atk | Privacy | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2022 | ACM | ETH | def | Privacy | Both | Other | ✓ | × | × | B | | × | L | L | 0 | 2 |
| 2022 | ACM | Hanqing | atk | Evasion | DL | Audio | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2022 | ACM | Stevens | atk | Privacy | DL | Other | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2022 | ACM | Are | atk | Privacy | DL | Other | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2022 | ACM | When | atk | Evasion | DL | Audio | ✓ | ✓ | ✓ | B | | × | H | H | 0 | 1 |
| 2022 | ACM | Learning | atk | Privacy | DL | Images | ✓ | ✓ | × | B | | × | H | L | 0 | 2 |
| 2022 | ACM | Using | atk | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2022 | ACM | Neural | atk | Evasion | DL | Text | ✓ | ✓ | × | B | P | ✓ | H | H | 0 | 3 |
| 2022 | ACM | CISPA | atk | Privacy | DL | Images | ✓ | ✓ | × | B | P | × | L | L | 0 | 1 |
| 2022 | ACM | A | atk | Privacy | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2022 | ACM | LPGNet: | def | Privacy | DL | Other | ✓ | ✓ | × | B | | ✓ | H | L | 0 | 3 |
| 2022 | IEEE | Encoders | atk | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2022 | IEEE | 1Illinois | atk | Evasion | DL | Images | ✓ | ✓ | × | B | | × | L | L | 0 | 1 |
| 2022 | IEEE | Graph | atk | Privacy | DL | Other | ✓ | ✓ | × | B | F | ✓ | H | L | 0 | 3 |
| 2022 | IEEE | Imperceptible | atk | Evasion | DL | Text | ✓ | ✓ | ✓ | B | | × | L | | 0 | 0 |
| 2022 | IEEE | Security | def | | DL | Other | ✓ | ✓ | × | B | | × | | L | 0 | 1 |
| 2022 | IEEE | Jialuo | def | | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2022 | IEEE | Yun | atk | Evasion | DL | Other | ✓ | ✓ | × | B | | × | H | L | 0 | 2 |
| 2022 | IEEE | Production | atk | Poisoning | DL | Images | ✓ | × | × | W | P | ✓ | H | L | 1 | 5 |
| 2022 | IEEE | Classification | def | | Both | Malware | ✓ | ✓ | × | B | | × | L | L | 0 | 1 |
| 2022 | IEEE | in | def | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2022 | IEEE | Markus | def | Privacy | Both | Other | ✓ | ✓ | × | B | | × | L | L | 0 | 1 |
| 2022 | IEEE | Eugene | atk | Poisoning | DL | Text | ✓ | ✓ | × | B | | ✓ | H | H | 0 | 3 |
| 2022 | IEEE | "Adversarial | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2022 | IEEE | Attacks | atk | Privacy | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2022 | IEEE | Weight | atk | Privacy | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2022 | IEEE | SoK: | def | | Both | Other | ✓ | × | × | B | | × | | L | 0 | 2 |
| 2022 | IEEE | Borja | atk | Privacy | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2022 | IEEE | Transformer | atk | Evasion | DL | Text | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2022 | IEEE | Property | atk | Poisoning | DL | Images | ✓ | ✓ | × | B | F | × | H | L | 0 | 2 |
| 2022 | IEEE | Yuhao | atk | Evasion | DL | Images | ✓ | ✓ | × | B | | × | H | L | 0 | 2 |
| 2022 | NDS | School | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2022 | NDS | Federated | def | Poisoning | DL | Text | ✓ | × | × | W | | × | L | L | 1 | 3 |
| 2022 | NDS | Property | atk | Privacy | DL | Images | ✓ | ✓ | × | B | F | ✓ | H | L | 0 | 3 |
| 2022 | NDS | RamBoAttack: | atk | Evasion | DL | Images | ✓ | ✓ | × | B | | × | H | H | 0 | 2 |
| 2022 | NDS | Shengwei | atk | Privacy | DL | Images | ✓ | ✓ | ✓ | W | F | ✓ | H | L | 1 | 3 |
| 2022 | NDS | Robustness | both | Poisoning | DL | Images | ✓ | × | × | W | F | ✓ | H | L | 1 | 5 |
| 2022 | NDS | Ahmed | atk | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2022 | USENIX | for | def | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2022 | USENIX | Vocal | def | Evasion | DL | Audio | ✓ | ✓ | × | B | | × | | L | 0 | 1 |
| 2022 | USENIX | 31st | atk | Privacy | Both | Other | ✓ | × | × | W | F | ✓ | H | L | 1 | 5 |
| 2022 | USENIX | August | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2022 | USENIX | sponsored | atk | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2022 | USENIX | August | atk | Privacy | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |

| Year | Venue | Paper (1st author) | G1 | G2 | G3 | G4 | G5 | G6 | G7 | T1 | T2 | Grad | Qry | Comp | WB | Trad. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2022 | USENIX | sponsored | atk | Privacy | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2022 | USENIX | sponsored | atk | Poisoning | DL | Text | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2022 | USENIX | Attacks | atk | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2022 | USENIX | sponsored | def | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2022 | USENIX | August | atk | Privacy | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2022 | USENIX | August | atk | Privacy | DL | Other | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2022 | USENIX | Novel | atk | Privacy | Both | Other | ✓ | ✓ | × | B | | × | H | L | 0 | 2 |
| 2022 | USENIX | Technische | def | | Both | Malware | ✓ | × | × | B | | × | | L | 0 | 2 |
| 2022 | USENIX | ICISPA | atk | Privacy | DL | Other | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | ACM | CISPA | def | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | ACM | Machine | def | Privacy | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | ACM | Virginia | atk | Poisoning | DL | Images | ✓ | ✓ | × | G | P | ✓ | L | L | 0 | 2 |
| 2023 | ACM | through | atk | Privacy | DL | Audio | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | ACM | Stealing | atk | Evasion | DL | Text | ✓ | ✓ | × | B | | × | H | L | 0 | 2 |
| 2023 | ACM | Stolen | both | Multiple | DL | Other | ✓ | ✓ | × | B | P | ✓ | H | L | 0 | 3 |
| 2023 | ACM | Information | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | ACM | Deep | def | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | ACM | Unforgeability | def | | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | ACM | Stefano | def | Evasion | Traditional | Images | ✓ | ✓ | × | B | | × | L | L | 0 | 1 |
| 2023 | ACM | Learning | atk | Poisoning | DL | Other | ✓ | ✓ | × | G | P | ✓ | H | L | 0 | 3 |
| 2023 | ACM | Secure | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | ACM | Detection | atk | Evasion | DL | Malware | ✓ | ✓ | × | B | | × | L | L | 0 | 1 |
| 2023 | ACM | Code | def | | DL | Other | ✓ | ✓ | × | G | | × | L | L | 0 | 1 |
| 2023 | ACM | and | def | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | ACM | through | atk | Privacy | DL | Other | ✓ | ✓ | × | B | | ✓ | H | L | 0 | 3 |
| 2023 | ACM | Jingxuan | def | | DL | Other | ✓ | ✓ | × | B | | × | L | L | 0 | 1 |
| 2023 | ACM | and | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | ACM | Evading | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | ACM | Changqing | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | ACM | AntiFake: | def | Evasion | DL | Audio | ✓ | × | ✓ | B | | × | H | L | 0 | 2 |
| 2023 | ACM | against | def | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | IEEE | Effects | def | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | IEEE | Deep | def | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | IEEE | arXiv:2112.04558v2 | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | IEEE | Federated | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | IEEE | SNAP: | atk | Poisoning | DL | Other | ✓ | ✓ | × | B | | × | H | L | 0 | 2 |
| 2023 | IEEE | Federated | def | Poisoning | DL | Images | ✓ | × | × | W | F | ✓ | H | L | 1 | 5 |
| 2023 | IEEE | Information | atk | Privacy | DL | Text | ✓ | ✓ | × | B | | × | H | L | 0 | 2 |
| 2023 | IEEE | Covert | atk | Poisoning | DL | Images | ✓ | ✓ | × | B | | ✓ | H | L | 0 | 3 |
| 2023 | IEEE | DepthFake | atk | Evasion | DL | Images | ✓ | ✓ | × | B | | × | H | L | 0 | 2 |
| 2023 | IEEE | Jiameng | atk | Evasion | DL | Text | ✓ | ✓ | × | B | | × | | L | 0 | 1 |
| 2023 | IEEE | Multi-party | def | Privacy | Both | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | IEEE | Andrew | def | Privacy | Both | Other | ✓ | × | × | W | F | ✓ | H | L | 1 | 5 |
| 2023 | IEEE | Information | atk | Privacy | DL | Text | ✓ | ✓ | × | B | | × | H | L | 0 | 2 |
| 2023 | IEEE | Extraction | atk | Evasion | DL | Images | ✓ | ✓ | × | B | | × | H | L | 0 | 2 |
| 2023 | IEEE | Limin | atk | Poisoning | DL | Malware | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | IEEE | Learning | def | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | IEEE | Attacks | both | Evasion | DL | Text | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | IEEE | Automatic | atk | Evasion | DL | Audio | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | IEEE | yCybersecurity | def | | DL | Images | ✓ | × | × | W | | × | L | L | 1 | 3 |
| 2023 | IEEE | Ruijie | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | IEEE | A | both | Privacy | DL | Images | ✓ | ✓ | × | B | P | × | H | L | 0 | 2 |
| 2023 | IEEE | SNAP: | atk | Poisoning | DL | Other | ✓ | × | × | B | | × | H | L | 0 | 3 |
| 2023 | IEEE | Hong | def | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | IEEE | BayBFed | def | Poisoning | DL | Images | ✓ | × | × | W | F | ✓ | H | L | 1 | 5 |
| 2023 | IEEE | Andre | atk | Evasion | DL | Audio | ✓ | ✓ | × | B | F | × | L | L | 0 | 1 |
| 2023 | IEEE | KASTEL | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | IEEE | ELSA: | def | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | IEEE | Shengwei | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | IEEE | Patch-agnostic | def | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | IEEE | RAB: | def | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | IEEE | ETH | def | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | IEEE | and | def | Privacy | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | IEEE | through | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | IEEE | StyleFool: | atk | Evasion | DL | Images | ✓ | ✓ | × | B | | × | L | L | 0 | 1 |
| 2023 | NDS | DNN | def | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | NDS | Hadi | def | Evasion | DL | Audio | ✓ | ✓ | × | B | | × | L | L | 0 | 2 |

| Year | Venue | Paper (1st author) | G1 | G2 | G3 | G4 | G5 | G6 | G7 | T1 | T2 | Grad | Qry | Comp | WB | Trad. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2023 | NDS | Yugeng | atk | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | NDS | Commercial | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | NDS | Classifiers | def | Evasion | DL | Images | ✓ | ✓ | × | B | | × | L | L | 0 | 1 |
| 2023 | NDS | Learning | def | Evasion | DL | Other | ✓ | ✓ | × | B | | ✓ | H | L | 0 | 3 |
| 2023 | NDS | Jiayun | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | NDS | Klim | atk | Evasion | DL | Other | ✓ | × | × | B | | × | H | H | 0 | 3 |
| 2023 | NDS | Siyuan | atk | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | NDS | Universidad | def | Poisoning | DL | Text | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | NDS | Deep | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | NDS | Chunyi | atk | Privacy | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | NDS | Robust | def | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | NDS | Physical | def | Evasion | Both | Other | ✓ | ✓ | ✓ | G | P | × | L | L | 0 | 0 |
| 2023 | NDS | *Technische | def | Privacy | Both | Text | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | USENIX | | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | USENIX | Detection | atk | Evasion | DL | Malware | ✓ | × | × | B | | × | H | L | 0 | 3 |
| 2023 | USENIX | Scientific | atk | Privacy | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | USENIX | of | def | | DL | Other | ✓ | ✓ | × | B | | × | L | L | 0 | 1 |
| 2023 | USENIX | for | atk | Poisoning | Both | Other | ✓ | × | × | W | F | ✓ | H | L | 1 | 5 |
| 2023 | USENIX | via | def | Privacy | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | USENIX | FREE | def | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | USENIX | 1CISPA | def | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | USENIX | Jonathan | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | USENIX | Vehicles | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | USENIX | Fairness | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | USENIX | Fairness | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | USENIX | Decompiling | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | USENIX | August | def | | Both | Malware | ✓ | ✓ | × | B | | × | L | L | 0 | 1 |
| 2023 | USENIX | Mazal | def | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | USENIX | August | def | | DL | Malware | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | USENIX | is | atk | Privacy | DL | Images | ✓ | ✓ | × | B | P | ✓ | H | L | 0 | 3 |
| 2023 | USENIX | is | def | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | USENIX | Federated | def | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | USENIX | Kunal | atk | Evasion | DL | Malware | ✓ | ✓ | × | B | P | × | H | L | 0 | 2 |
| 2023 | USENIX | for | def | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | USENIX | by | def | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | USENIX | is | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | USENIX | is | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | USENIX | August | atk | Poisoning | DL | Text | ✓ | × | × | B | | × | H | L | 0 | 3 |
| 2023 | USENIX | Universal | atk | Evasion | DL | Images | ✓ | ✓ | × | B | | × | H | L | 0 | 2 |
| 2023 | USENIX | is | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2023 | USENIX | August | def | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | ACM | from | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | ACM | Xueluan | def | Privacy | DL | Images | ✓ | × | × | B | | × | L | L | 0 | 2 |
| 2024 | ACM | SafeEar: | def | Privacy | DL | Audio | ✓ | ✓ | × | B | | × | L | L | 0 | 1 |
| 2024 | ACM | Against | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | ACM | Multi-Agent | atk | Evasion | DL | Images | ✓ | ✓ | × | B | | × | H | H | 0 | 2 |
| 2024 | ACM | Text-to-Image | def | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | ACM | Autonomous | def | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | ACM | Wei | atk | Poisoning | Both | Other | ✓ | ✓ | × | G | P | ✓ | H | L | 0 | 3 |
| 2024 | ACM | Byzantine-Robust | def | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | ACM | Recognition | atk | Evasion | DL | Audio | ✓ | ✓ | ✓ | B | | × | | H | 0 | 0 |
| 2024 | ACM | Zijin | def | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | ACM | CSIRO's | atk | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | ACM | Items | atk | Poisoning | DL | Other | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | ACM | Applications | atk | Privacy | DL | Text | ✓ | ✓ | × | B | | ✓ | H | H | 0 | 3 |
| 2024 | ACM | Attack | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | ACM | Attachment | atk | Evasion | DL | Malware | ✓ | ✓ | ✓ | B | | × | H | L | 0 | 1 |
| 2024 | ACM | Information | atk | Evasion | DL | Text | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | ACM | Alchemy: | def | Evasion | DL | Images | ✓ | ✓ | × | W | | ✓ | H | L | 1 | 4 |
| 2024 | ACM | S2NeRF: | both | Privacy | DL | Images | ✓ | ✓ | × | B | | ✓ | H | L | 0 | 3 |
| 2024 | ACM | Certified | atk | Poisoning | DL | Other | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | ACM | ETH | atk | Privacy | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | ACM | Jie | atk | Privacy | DL | Images | ✓ | ✓ | × | B | P | × | H | L | 0 | 2 |
| 2024 | ACM | Understanding | atk | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | ACM | in | def | Evasion | DL | Malware | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | ACM | A | def | Privacy | DL | Images | ✓ | ✓ | × | B | | × | H | L | 0 | 2 |

| Year | Venue | Paper (1st author) | G1 | G2 | G3 | G4 | G5 | G6 | G7 | T1 | T2 | Grad | Qry | Comp | WB | Trad. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2024 | ACM | The | def | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | ACM | Fisher | def | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | ACM | into | atk | Privacy | DL | Text | ✓ | ✓ | × | B | | × | H | L | 0 | 2 |
| 2024 | ACM | BadMerging: | atk | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | ACM | Optimization-based | atk | Evasion | DL | Text | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | ACM | Optimization-based | atk | Evasion | DL | Text | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | ACM | Model | atk | Privacy | DL | Text | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | ACM | Shenzhen | atk | Privacy | DL | Text | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | ACM | Information | def | Evasion | DL | Images | ✓ | ✓ | × | B | | × | | H | 0 | 1 |
| 2024 | ACM | Learning | atk | Privacy | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | ACM | Blind | def | | Both | Audio | × | × | × | B | | × | | L | 0 | 3 |
| 2024 | ACM | Information | atk | Privacy | DL | Text | × | × | × | B | | × | L | L | 0 | 2 |
| 2024 | ACM | Deepfake | def | | DL | Audio | ✓ | ✓ | × | B | | × | | L | 0 | 1 |
| 2024 | ACM | Membership | atk | Privacy | DL | Images | ✓ | ✓ | × | B | P | × | L | L | 0 | 1 |
| 2024 | ACM | Membership | atk | Privacy | DL | Images | ✓ | ✓ | × | B | P | × | L | L | 0 | 1 |
| 2024 | ACM | Neural | atk | Privacy | DL | Images | ✓ | ✓ | × | B | | × | | L | 0 | 1 |
| 2024 | ACM | | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | ACM | Fine-grained | def | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | ACM | from | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | ACM | Institute | atk | Privacy | DL | Images | ✓ | ✓ | × | B | | × | H | L | 0 | 2 |
| 2024 | ACM | Institute | atk | Privacy | DL | Images | ✓ | ✓ | × | B | | × | H | L | 0 | 2 |
| 2024 | ACM | of | def | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | ACM | School | atk | Evasion | DL | Text | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | ACM | Examples: | atk | Evasion | DL | Images | ✓ | ✓ | × | B | | × | H | L | 0 | 2 |
| 2024 | ACM | Models | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | ACM | Learning | atk | Poisoning | DL | Images | ✓ | ✓ | × | B | P | × | L | L | 0 | 1 |
| 2024 | IEEE | Pattern | def | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | IEEE | Robin | def | Privacy | Both | Other | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | IEEE | Yuzheng | def | Privacy | Both | Images | ✓ | ✓ | × | B | | × | L | L | 0 | 1 |
| 2024 | IEEE | Joshua | atk | Privacy | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | IEEE | Ziqi | def | Privacy | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | IEEE | Linguistic | atk | Evasion | DL | Audio | ✓ | ✓ | ✓ | B | | × | L | L | 0 | 0 |
| 2024 | IEEE | Language | atk | Evasion | DL | Text | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | IEEE | It's | def | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | IEEE | into | atk | Evasion | DL | Images | ✓ | ✓ | × | B | | × | H | L | 0 | 2 |
| 2024 | IEEE | Test-Time | atk | Poisoning | DL | Images | ✓ | ✓ | × | G | P | ✓ | L | L | 0 | 2 |
| 2024 | IEEE | Xingshuo | atk | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | IEEE | Mohammad | atk | Poisoning | DL | Images | ✓ | × | × | G | P | ✓ | H | L | 0 | 4 |
| 2024 | IEEE | Attacks | def | Evasion | DL | Images | ✓ | ✓ | × | B | | × | H | L | 0 | 2 |
| 2024 | IEEE | Distribution | atk | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | IEEE | Dropout | atk | Poisoning | DL | Images | ✓ | × | × | W | F | ✓ | H | L | 1 | 5 |
| 2024 | IEEE | GROV | def | Evasion | DL | Other | ✓ | ✓ | × | B | | × | L | L | 0 | 1 |
| 2024 | IEEE | Mahmoud | atk | Evasion | DL | Other | ✓ | ✓ | × | B | P | × | L | L | 0 | 1 |
| 2024 | IEEE | Alec | def | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | IEEE | Accelerator | atk | Privacy | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | IEEE | KASTEL | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | IEEE | Adversarial | def | Evasion | DL | Text | × | ✓ | × | W | | × | | L | 1 | 3 |
| 2024 | IEEE | 2Services | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | NDS | Large | def | Evasion | DL | Text | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | NDS | DEMASQ: | def | Evasion | DL | Text | ✓ | × | × | B | | ✓ | H | L | 0 | 4 |
| 2024 | NDS | Attacks | atk | Privacy | DL | Audio | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | NDS | Poisoning | atk | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | NDS | by | def | Privacy | DL | Images | ✓ | ✓ | × | B | | × | L | L | 0 | 1 |
| 2024 | NDS | CamPro | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | NDS | Ensemble | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | NDS | Gelei | atk | Evasion | DL | Text | ✓ | × | × | B | | × | H | H | 0 | 3 |
| 2024 | NDS | Watermarking | def | Evasion | DL | Audio | ✓ | ✓ | ✓ | W | F | × | L | L | 1 | 1 |
| 2024 | NDS | ICS | atk | Evasion | DL | Other | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | NDS | Federated | def | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | NDS | Chengkun | def | Poisoning | DL | Text | ✓ | ✓ | × | B | | ✓ | L | L | 0 | 2 |
| 2024 | NDS | Recognition | atk | Evasion | DL | Audio | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | NDS | Transpose | atk | Privacy | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | NDS | Transpose | atk | Privacy | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | NDS | Inversion-based | def | Privacy | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |

| Year | Venue | Paper (1st author) | G1 | G2 | G3 | G4 | G5 | G6 | G7 | T1 | T2 | Grad | Qry | Comp | WB | Trad. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2024 | NDS | Networks | atk | Poisoning | DL | Images | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | NDS | Browser | def | Privacy | Traditional | Other | √ | × | × | B |  | × |  | L | 0 | 2 |
| 2024 | NDS | Deep | def | Evasion | DL | Images | √ | √ | × | B |  | × | H | L | 0 | 2 |
| 2024 | NDS | Bang | def | Privacy | DL | Other | √ | √ | × | B |  | × | L | L | 0 | 1 |
| 2024 | NDS | Against | atk | Poisoning | DL | Images | √ | √ | × | B | P | √ | H | L | 0 | 3 |
| 2024 | NDS | Speaker | atk | Evasion | DL | Audio | √ | √ | √ | B |  | × |  | L | 0 | 0 |
| 2024 | NDS | Mitigating | def | Poisoning | DL | Images | √ | × | × | B |  | × |  | L | 0 | 2 |
| 2024 | NDS | for | def |  | DL | Malware | √ | √ | × | B |  | × | L | L | 0 | 1 |
| 2024 | NDS | in | both | Evasion | DL | Images | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | NDS | Attacks | def | Poisoning | DL | Text | √ | √ | × | W | F | × | L | L | 1 | 2 |
| 2024 | NDS | Chaoxiang | atk | Evasion | DL | Images | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | USENIX | Tong | atk | Evasion | DL | Text | √ | √ | × | B |  | × | L | L | 0 | 1 |
| 2024 | USENIX | Attacks | atk | Evasion | DL | Images | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | USENIX | Approach | def | Poisoning | DL | Images | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | USENIX | Qingzhao | atk | Evasion | DL | Images | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | USENIX | Zhenghang | def |  | Both | Other | √ | √ | × | B |  | × |  | L | 0 | 1 |
| 2024 | USENIX | Synthesis | def | Privacy | DL | Images | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | USENIX | Measuring | def |  | DL | Audio | √ | √ | × | B |  | × | L | L | 0 | 1 |
| 2024 | USENIX | Large | atk | Evasion | DL | Text | √ | √ | × | B |  | × | H | L | 0 | 2 |
| 2024 | USENIX | Efficiency | def | Evasion | DL | Images | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | USENIX | Efficiency | def | Evasion | DL | Images | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | USENIX | Kathrin | atk | Evasion | DL | Images | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | USENIX | DNN-GP: | both | Evasion | DL | Images | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | USENIX | for | atk | Evasion | DL | Images | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | USENIX | DeepEclipse | atk | Evasion | DL | Images | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | USENIX | DeepEclipse | atk | Evasion | DL | Images | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | USENIX | Between | atk | Multiple | DL | Images | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | USENIX | Red-Teaming | atk | Evasion | DL | Images | √ | √ | × | B |  | × | H | L | 0 | 2 |
| 2024 | USENIX | Red-Teaming | atk | Evasion | DL | Images | √ | √ | × | B |  | × | H | L | 0 | 2 |
| 2024 | USENIX | Property | atk | Privacy | DL | Images | √ | √ | × | B | F | × | H | L | 0 | 2 |
| 2024 | USENIX | Transposed | def |  | DL | Images | √ | × | × | W | F | √ | H | L | 1 | 5 |
| 2024 | USENIX | in | def | Privacy | DL | Images | √ | √ | × | B |  | × | L | L | 0 | 1 |
| 2024 | USENIX | Shuaifan | def | Privacy | DL | Images | √ | √ | √ | W | F | √ | H | L | 1 | 3 |
| 2024 | USENIX | CISPA | atk | Evasion | DL | Images | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | USENIX | CISPA | atk | Evasion | DL | Images | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | USENIX | Runtime | atk | Evasion | DL | Images | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | USENIX | Splitting | def | Evasion | DL | Images | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | USENIX | Yixin | atk | Privacy | DL | Images | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | USENIX | Changjiang | atk | Poisoning | DL | Images | √ | √ | × | B | F | √ | H | L | 0 | 3 |
| 2024 | USENIX | Guangsheng | atk | Privacy | DL | Images | √ | × | × | W | F | √ | H | L | 1 | 5 |
| 2024 | USENIX | Bailey | def | Privacy | DL | Images | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | USENIX | Meng | atk | Evasion | DL | Images | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | USENIX | "Minxue | def | Privacy | DL | Images | √ | √ | × | W |  | √ | H | L | 1 | 4 |
| 2024 | USENIX | CISPA | atk | Privacy | DL | Images | √ | √ | × | B | P | × | H | H | 0 | 2 |
| 2024 | USENIX | False | atk | Evasion | DL | Images | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | USENIX | Shenchen | atk | Evasion | DL | Images | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | USENIX | through | atk | Privacy | DL | Other | √ | × | × | B |  | √ | H | L | 0 | 4 |
| 2024 | USENIX | Shaofeng | atk | Evasion | DL | Images | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | USENIX | Mudjacking: | def | Poisoning | DL | Images | √ | × | × | B |  | √ | H | L | 0 | 4 |
| 2024 | USENIX | Randomised | atk | Evasion | DL | Images | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | USENIX | 33rd | atk | Evasion | DL | Images | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | USENIX | is | def |  | DL | Images | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | USENIX | August | def | Privacy | Both | Text | √ | √ | × | B |  | × | L | L | 0 | 1 |
| 2024 | USENIX | August | atk | Privacy | DL | Images | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | USENIX | is | atk | Poisoning | DL | Images | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | USENIX | is | def |  | DL | Images | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | USENIX | is | def | Poisoning | DL | Images | √ | × | × | B |  | × | L | L | 0 | 2 |
| 2024 | USENIX | Attacks | def | Privacy | DL | Images | √ | √ | × | B |  | √ | H | L | 0 | 3 |
| 2024 | USENIX | Model | atk | Evasion | DL | Text | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | USENIX | Zooming | both | Evasion | DL | Images | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | USENIX | August | atk | Privacy | DL | Text | √ | √ | × | B |  | × | L | L | 0 | 1 |
| 2024 | USENIX | is | def |  | DL | Text | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | USENIX | August | def |  | DL | Other | √ | √ | × | W | F | × | H | L | 1 | 3 |
| 2024 | USENIX | August | def |  | DL | Other | √ | √ | × | W | F | √ | H | L | 1 | 4 |
| 2024 | USENIX | August | def | Evasion | DL | Images | √ | √ | × | B |  | × | L | L | 0 | 1 |
| 2024 | USENIX | August | def |  | Both | Malware | √ | × | × | B |  | × | L | L | 0 | 2 |
| 2024 | USENIX | Learning | def | Evasion | Both | Other | √ | √ | × | W | F | √ | H | L | 1 | 4 |

39

| Year | Venue | Paper (1st author) | G1 | G2 | G3 | G4 | G5 | G6 | G7 | T1 | T2 | Grad | Qry | Comp | WB | Trad. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2024 | USENIX | Framework | def | | DL | Text | ✓ | ✓ | × | B | | × | | L | 0 | 1 |
| 2024 | USENIX | | def | | DL | Images | ✓ | ✓ | × | B | | × | | L | 0 | 1 |
| 2024 | USENIX | | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2024 | USENIX | Inference | def | Privacy | DL | Images | ✓ | × | × | G | | ✓ | H | L | 0 | 4 |
| 2025 | ACM | 1Technion | atk | Evasion | DL | Text | ✓ | ✓ | × | B | | × | H | L | 0 | 2 |
| 2025 | ACM | in | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2025 | ACM | arXiv:2501.05928v2 | atk | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2025 | ACM | Anti-Facial | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2025 | ACM | BIFOLD | atk | Evasion | DL | Other | ✓ | ✓ | × | W | F | ✓ | H | H | 1 | 4 |
| 2025 | ACM | Targeted | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2025 | ACM | Language | atk | Poisoning | DL | Text | ✓ | ✓ | × | B | | × | L | L | 0 | 1 |
| 2025 | ACM | Models | def | Evasion | DL | Images | ✓ | × | ✓ | G | P | ✓ | H | H | 0 | 3 |
| 2025 | ACM | Adversarial | atk | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2025 | ACM | Construction | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2025 | ACM | Latent-based | atk | Evasion | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2025 | ACM | Milad | atk | Evasion | DL | Malware | ✓ | ✓ | ✓ | W | F | ✓ | H | L | 1 | 3 |
| 2025 | ACM | VillainNet: | atk | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2025 | ACM | Deep | def | | DL | Malware | ✓ | ✓ | × | B | | × | | L | 0 | 1 |
| 2025 | USENIX | Jiachen | def | Evasion | DL | Images | ✓ | ✓ | × | B | | ✓ | H | H | 0 | 2 |
| 2025 | USENIX | Lingchen | def | Evasion | DL | Text | ✓ | ✓ | × | W | F | × | L | L | 1 | 2 |
| 2025 | USENIX | Revisiting | atk | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2025 | USENIX | Zhisheng | def | Privacy | DL | Audio | ✓ | ✓ | × | B | | × | L | L | 0 | 1 |
| 2025 | USENIX | with | def | Evasion | DL | Other | ✓ | ✓ | × | W | F | × | H | L | 1 | 3 |
| 2025 | USENIX | Persistent | atk | Poisoning | DL | Images | ✓ | ✓ | × | W | | ✓ | H | L | 1 | 4 |
| 2025 | USENIX | Benign | atk | Poisoning | DL | Images | ✓ | ✓ | × | W | F | ✓ | H | L | 1 | 4 |
| 2025 | USENIX | 34th | def | Evasion | DL | Malware | ✓ | ✓ | × | B | | ✓ | H | L | 0 | 3 |