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 **4**

2 Background: Foundations of Adversarial Machine Learning **7**

2.1 The Discovery of Adversarial Vulnerabilities 7

2.2 The Arms Race: Attacks and Defenses 8

2.3 Expanding Threat Landscape 8

2.4 The Theory-Practice Gap 8

2.5 Types of Adversarial Attacks 9

2.6 Threat Models and Their Implications 9

3 Methodology **10**

3.1 Data Collection 10

3.2 Coding Framework 10

3.3 Gap Score Framework 11

4 Quantitative Findings **11**

4.1 Real-World Validation 11

4.2 Gradient Dependency 12

4.3 Threat Model Assumptions 13

4.4 Query Budget Assumptions 14

4.5 Domain Distribution 15

4.6 Code Availability 15

4.7 Gap Score Distribution 16

5 Thematic Findings **16**

5.1 The Utility-Robustness Trade-off 16

5.2 Evaluation Methodology 17

5.2.1 Distance Metrics 17

5.2.2 Privacy Evaluation 17

5.3 Physical Deployment Considerations 17

5.3.1 Physical Attack Challenges 17

5.3.2 System Integration 18

5.4 Attack Evolution 18

5.5 Defense Evolution 19

5.6 Economic and Organizational Factors 19

5.6.1 Economic Considerations 19

5.6.2 Organizational Considerations 19

6 Temporal Trends **20**

6.1 Stable Patterns 20

6.2 Areas of Modest Progress 20

6.3 Emerging Research Areas 21

7 Cross-Venue Analysis **21**
7.1 ACM CCS: Economics and System Integration . . . 21
7.2 IEEE S&P: Efficiency and Formal Guarantees . . . 22
7.3 NDSS: Physical Constraints and Operational Realities . . . 22
7.4 USENIX Security: Real-World Validation and Domain Diversity . . . 22
7.5 Attack versus Defense Papers . . . 22

8 Recommendations **23**
8.1 For Researchers . . . 23
8.1.1 Specify Realistic Threat Models . . . 23
8.1.2 Employ Realistic Evaluation Frameworks . . . 23
8.1.3 Report Practical Attack Economics . . . 24
8.2 For Conferences and Reviewers . . . 24
8.2.1 Require Artifact Availability . . . 24
8.2.2 Evaluate Threat Model Realism . . . 24
8.3 For Practitioners . . . 25
8.3.1 Implement Defense in Depth . . . 25
8.3.2 Conduct Continuous Adversarial Testing . . . 25
8.3.3 Contribute to Shared Threat Intelligence . . . 25
8.4 For Funders and Policymakers . . . 25
8.4.1 Require Deployment Validation . . . 25
8.4.2 Support Evaluation Infrastructure . . . 26
8.4.3 Address Critical Gaps . . . 26
8.4.4 Align Regulation with Research . . . 26
8.5 Path Forward . . . 26

9 Limitations and Threats to Validity **27**
9.1 Scope Limitations . . . 27
9.2 Coding Limitations . . . 27
9.3 Generalizability . . . 27

10 Conclusion **27**

A Complete Paper Analysis Dataset **38**

1 Introduction

Machine learning is becoming an important part of modern software systems. Applications range from face recognition systems at the border control entrance [1], fraud analysis in the financial industry [2], routing in self-driving cars [3], to moderation in social media sites [4]. Consequently, ML models affect decisions in systems that impact millions of users every day. This widespread deployment comes with the need for security and robustness against adversarial vulnerabilities, described in this paper as input or interaction crafted to cause models to behave in unintended and harmful ways [5].

Adversarial machine learning (AML) studies attacks on machine learning systems and defences that aim to improve their security and strength [6]. In the last decade, researchers have shown various attacks within the machine learning pipeline. These include evasion attacks, which create harmful inputs that lead to misclassification during inference [7, 8], poisoning attacks, which insert flawed training data to disrupt how models behave [9, 10], and privacy attacks, which pull out sensitive information through model queries or inspections [11, 12]. In response to these attacks, a substantial body of studies has proposed defences designed to enhance robustness, privacy, and resistance to adversarial influence [13, 14].

Real World Scenarios

Despite significant progress in research and theory being made, the consequences of adversarial vulnerabilities extend beyond academic benchmarks. For instance, in November 2025, Anthropic disclosed what it believed with high confidence to be the first large-scale state-sponsored cyber espionage campaign mostly carried out by an AI agent [15]. Attackers manipulated Claude Code to autonomously conduct reconnaissance, generate exploits, and exfiltrate data from around thirty big global targets. The system handled about 80–90% of the operational workload with minimal human oversight. Notably, this incident did not depend on gradient-based attacks, sophisticated mathematics or white-box model access. Instead, it exploited deployment-level assumptions about agentic autonomy, tool access via the Model Context Protocol [16], and circumvented safeguards through chunking cyberattacks into small "innocent" looking tasks [17]. This case illustrates a broader trend detailed in Section 3: real adversaries target system integration points, operational weaknesses, human workflows, and economic leverage rather than computing gradients or optimal L_p-bounded perturbations [18, 19].

The economic and social stakes are significant. Adversarial attacks lead to direct financial losses. For instance, deepfake-enabled fraud has resulted in documented losses of millions of dollars in individual cases [20]. Additionally, data extraction attacks can reveal valuable proprietary training data, leaking sensitive information or private user data [12]. Organisations struggle to implement defences against adversarial attacks due to unclear ownership of machine learning security risks, a lack of tools for adversarial testing, and a limited understanding of realistic threat models [21, 22]. Industry surveys show that many are unsure how to evaluate adversarial risks in operational systems [23]. At the same time, regulatory frameworks are increasingly requiring robustness guarantees for high-stakes machine learning implementations [24]. This creates compliance costs and possible penalties. However, based on recent surveys with industry, practitioners often find that academic defences do not work well under operational constraints [25].

The Apruzzese Critique

Prior work has revealed this gap. In 2022, Apruzzese et al published "Real Attackers Don't Compute Gradients" [26]. This analysis reveals a serious gap between academic research in adversarial machine

learning (AML) and real-world security issues. They found that much of the existing literature assumes attackers have abilities that are rarely available in practice. These include white-box access, meaning complete knowledge of model architectures, parameters and training methods [27], the ability to compute gradients through target models [28], virtually unlimited query budgets [29], and no monitoring or rate limiting [30].

Through studies of actual machine learning attacks, they showed that real attackers are driven by economic incentives and usually use simpler methods. Since attackers have limited knowledge, budget and computation power, they often rely on basic input manipulation, exploiting weaknesses in the system, or social engineering when these tactics are cheap enough to meet their profitable goals. This critique raises an important question about whether academic research priorities match operational security needs.

Research Questions

Building upon this work, we conduct a systematic review to answer the following questions:

- **RQ1:** To what extent has Adversarial ML research addressed practical deployment constraints in recent years (2022–2025)?
- **RQ2:** What patterns distinguish research that bridges the theory-practice gap from purely laboratory-focused work?
- **RQ3:** How is this gap relevant to emerging threats in large language models and foundational models? How should future research address these new paradigms?

Positioning in Prior Work

This review builds on a growing body of meta-analyses evaluating the practical security research. Kumar et al [22] conducted interviews with 28 organisations and found that practitioners lack relevant tools and understanding for deploying robust machine learning systems. Grosse et al [23] surveyed 271 industry practitioners and revealed systematic overestimation of realistic attacker capabilities by academic work. Arp et al [25] performed a meta-analysis of security research and documented reliance on simplified laboratory setups that don't account for adaptive adversaries. Apruzzese et al. [26] justified these concerns through real-world attack case studies and revealed that real ML attacks rarely resemble academic threat models.

Our work expands on this topic in four ways. First, we offer the first systematic analysis across different venues during the post-Apruzzese period (2022–2025). We look at how recent research compares to the prior work. Second, we introduce a quantitative Gap Score framework that we apply consistently to 454 papers. This allows for statistical comparisons of research practices across different venues and over time. Third, we note the rise of new threat landscapes, especially agentic LLM misuse and multimodal attacks. These present challenges that differ from traditional computer vision issues. Fourth, we provide specific recommendations for four groups of stakeholders instead of general observations.

Scope and Boundaries

To address these questions, we analyse 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 chose these venues because they are top security research outlets that emphasise real-world impact and practical relevance in their calls for papers and review criteria. The 2022 to 2025 timeframe captures the research community's response after critiques, giving us enough coverage to identify trends while keeping methodical consistency.

Our analysis does not include papers from general ML conferences (NeurIPS, ICML, ICLR), workshop-only publications, and preprints. This choice is intentional. Security-focused venues have different review criteria and author motivations than ML theory venues, where improving benchmark performance may take precedence over deployment concerns. Security conference review criteria stress realism in threat models, evaluation on operational systems, and consideration of attacker motivations. By focusing on security conferences, we look at research that appears to prioritise practical security impact, highlighting the theory–practice gap when it exists.

Each paper is evaluated using dimensions adapted from Apruzzese et al. [26]. We consider threat model assumptions (white-box, gray-box, or black-box adversary knowledge [27]), reliance on gradient access (whether attacks require backpropagation through the target model), query budget needs (number of model inferences required), computational cost (GPU needs and training time), validation setting (evaluation on real production systems versus static benchmarks), and economic or organizational constraints (costs, incentives, and operational barriers). These dimensions are turned into a simple Gap Score (0 to 6), where 0 means full alignment with practical constraints, and 6 means reliance on all six idealised assumptions (detailed in Section 4).

Preview of Key Findings

Our analysis shows that the theory–practice gap identified in 2022 is still significant. Key findings include:

- *Limited real-world evaluation:* Only 5.3% of papers (24 out of 454) evaluate attacks or defences in deployed systems (operational environments with real users, monitoring, and operational constraints). The rest rely on offline datasets and simulated environments.
- *Persistent gradient dependence:* 67.8% of papers need gradient information from target models (access to model internals that enable backpropagation-based attacks), with a slight change from 2022 (68.2%) to early 2025 (65.9%).
- *White-box dominance:* 63.2% assume white-box adversaries with complete model knowledge, even though industry surveys show that such access is rare. In contrast, black-box setups (query-only access without model internals) better reflect most deployment situations.
- *High query budgets:* Among papers involving model queries, 80.4% assume budgets over 1000 queries, often overlooking cost limits, rate limiting, and anomaly detection systems.
- *Average Gap Score of 3.17:* Most papers use about half of the idealised assumptions we track, with only 10.5% scoring 0 to 1, indicating close alignment with deployment constraints.

Despite these gaps, we see some positive trends. Code release rates have gone up, improving reproducibility. New studies are looking into query-efficient and gradient-free methods. There is increasing focus on threats specific to LLMs, like jailbreaking and prompt injection, along with some movement beyond computer vision into text, audio, and malware detection.

Contributions

This review makes the following contributions:

- *First post-Apruzzese cross-venue systematic analysis:* We offer the first comprehensive assessment of how 454 papers published across four leading security venues (ACM CCS, IEEE S&P, NDSS, USENIX Security, 2022–2025) deal with the theory–practice gap in recent years.

- *Quantitative Gap Score framework:* We introduce a simple coding framework with a six-dimensional Gap Score, allowing for statistical comparison of practical relevance across venues, years, and research focuses (attack versus defence). This extends earlier qualitative meta-analyses.
- *Documentation of recent real-world adversarial incidents:* We summarise documented cases of ML attacks in deployed systems, including the first large-scale agentic AI cyber campaign, prompt injection vulnerabilities, and deepfake fraud, showing the sociotechnical nature of recent adversarial threats.
- *Emergence analysis of the LLM threat landscape:* We describe how foundation models and agentic systems introduce new adversarial challenges like jailbreaking, prompt injection, and multimodal attacks that differ from traditional threat models.
- *Actionable stakeholder-specific recommendations:* We offer targeted guidance for researchers (evaluation methods), conference organisers (review criteria), practitioners (deployment considerations), and funders (bridging infrastructure), based on our findings.

Who This Review Serves

This work serves several groups. This can help researchers who design new attacks/defences or benchmarks use our finding to make evaluation methods fit real world situations. Conference program committees and reviewers can refer to our quantitative analysis and improvise a criterion that further prioritises real-world constraints. Industry practitioners can use our gap model to assess whether proposed attacks/defences are relevant in the real world. This helps them avoid using methods that only work under ideal situations. Funding agencies can take into consideration our findings and encourage collaboration between academia and industry, set realistic goals, and build effective infrastructure. Overall, we hope this work inspires a growing body of adversarial machine learning research to evaluate their work for real-world scenarios and bridge the gap between theory and practice.

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. [31]. 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. [32] 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. [33] 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_∞), achieving high success rates against certain defenses, including defensive distillation [35].

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 [36] 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. [37] introduced ensemble adversarial training to enhance black-box robustness. Research on certified robustness [38] 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, 39] determine whether specific data points were part of model training. Carlini et al. [40] 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 [41], along with distribution-preserving methods that avoid detection [42]. Poisoning attacks that corrupt training data demonstrate that even small amounts of malicious samples can significantly harm model performance [43].

Physical-world attacks took gradient-based techniques beyond digital settings. Kurakin et al. [44] showed that adversarial examples remained effective when printed and photographed. Later work [45] 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 or-

ganizations, discovering that the industry lacked practical tools for dealing with adversarial ML threats. Grosse et al. [46] 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.

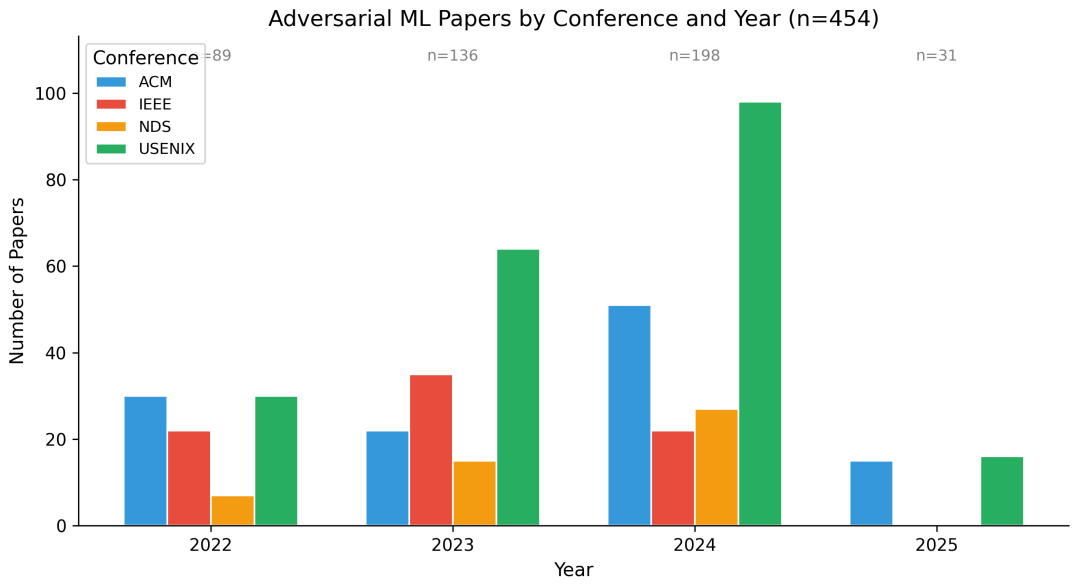

Figure 1: Dataset overview showing the distribution of 454 papers across four security conferences (ACM CCS, IEEE S&P, NDSS, USENIX Security) from 2022 to 2025.

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

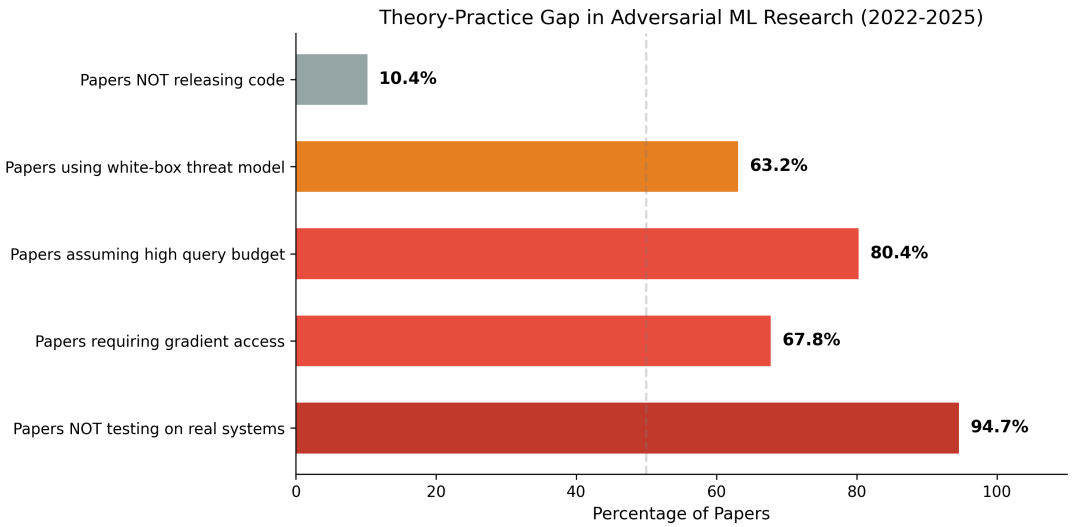

Figure 2: Overview of the theory-practice gap showing the percentage of papers exhibiting each characteristic: no real-world testing (94.7%), high query budget (80.4%), gradient dependency (67.8%), white-box access (63.2%), and no code release (10.4%).

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.

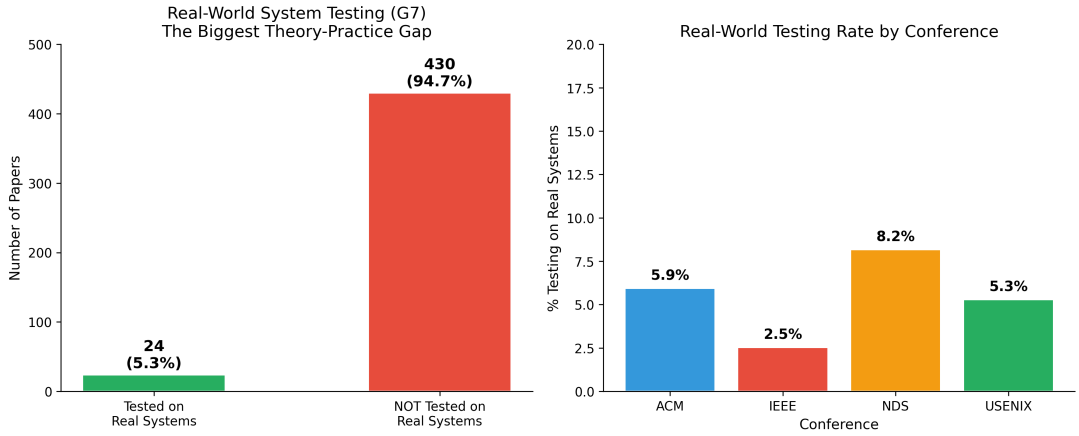

Figure 3: Real-world testing rate: 5.3% of papers validate on deployed systems.

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. [47] 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. [48] demonstrated that deepfake detection research employs metrics and dataset distributions that may lead to overestimation of detector efficacy.

Duan et al. [49] 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.

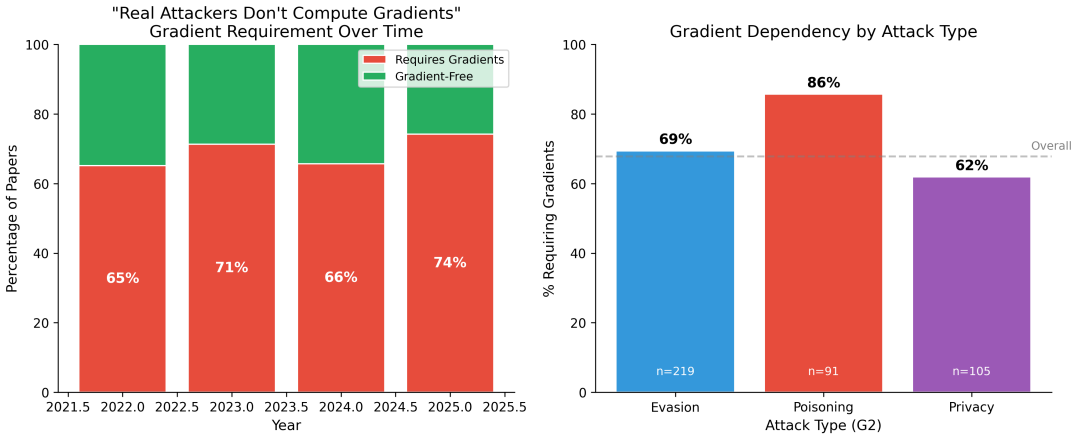

Figure 4: Gradient dependency analysis showing limited change over time (68.2% in 2022 to 65.9% in 2025) and variation by attack type.

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. [50] 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. [51] and Yu et al. [52] 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.

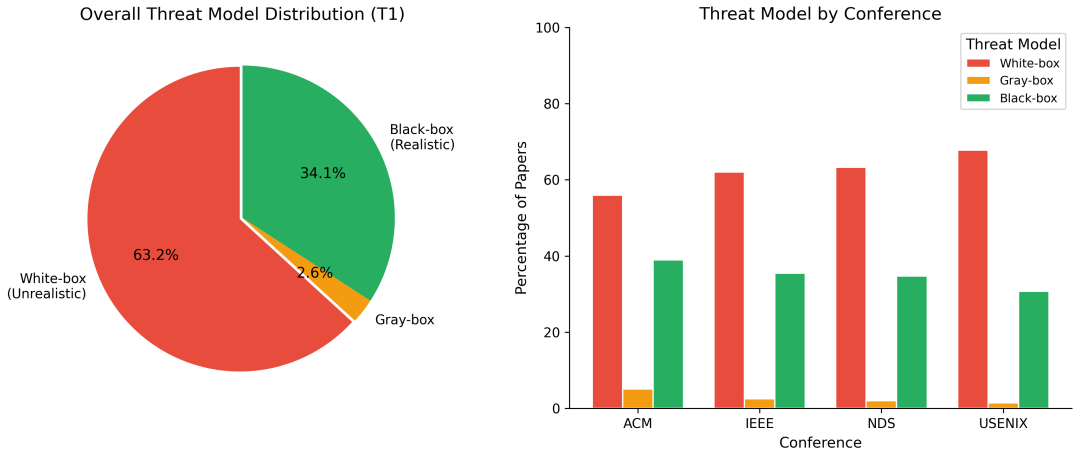

Figure 5: Distribution of threat model assumptions across all papers, showing prevalence of white-box assumptions.

This pattern contrasts with industrial practice. As Grosse et al. [46] 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. [53] 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.

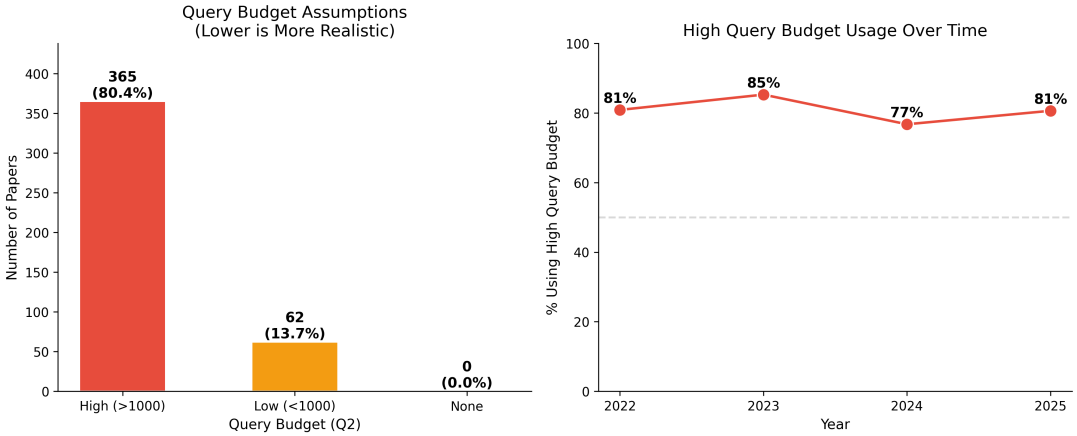

Figure 6: Query budget assumptions showing 80.4% of papers assume high query budgets.

Some recent work has addressed query efficiency. HARDBEAT from Tao et al. [54] 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. [55] 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%.

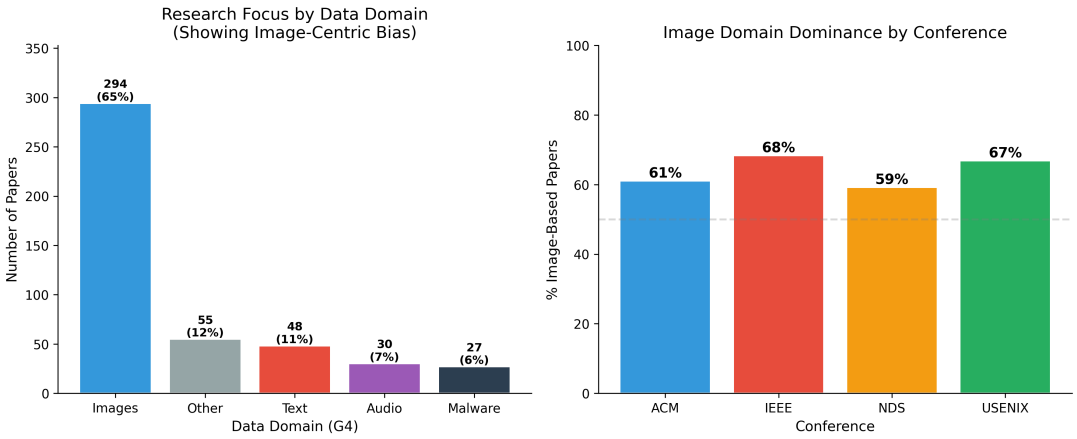

Figure 7: Data domain distribution revealing concentration on image classification in adversarial ML research.

This concentration creates potential blind spots. Financial systems process tabular data where adversarial perturbations cannot be measured by pixel distances. As Kireev et al. [56] 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

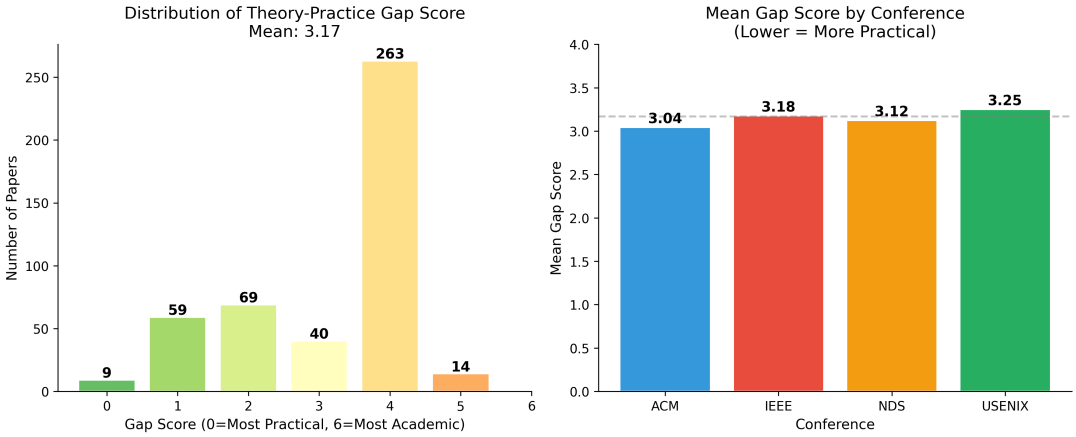

Figure 8: Distribution of Gap Scores showing the typical paper (mean 3.17/6) incorporates roughly half of the measured assumptions.

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. [57] 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. [58] 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. [59] 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 [60, 61].

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 [57]. MIST offers a pathway toward robust security by strategically limiting overfitting only to the most membership-vulnerable training instances [62]. 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 [63].

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. [40] 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 [64], 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. [40] 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. [45] 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. [65] 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. [66] demonstrated Physical Removal Attacks using focused laser spoofing to selectively remove LiDAR point cloud data on autonomous vehicles. The "Tubes Among Us" research [67] 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" [68] exploits LiDAR weaknesses using ordinary non-reflective materials placed on roads, requiring no specialized equipment. ATKSCOPES [69] 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. [70] 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. [71] 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. [72] 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.

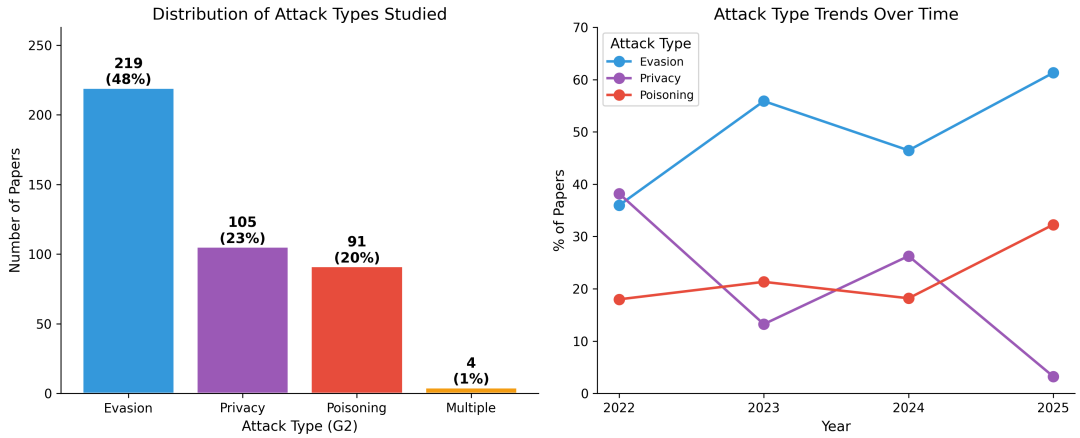

Figure 9: Attack type distribution showing evasion attacks (48%) predominate, with trends over time.

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 [73].

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 [74] reveals supply chain considerations where seemingly benign upstream models pass security checks but activate malicious backdoors upon merging with other components. Guo et al. [75] 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 [76] 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 [77] moves beyond heuristics for LLM protection by using the Linear Representation Hypothesis to identify and manipulate "toxic" and "jailbreak" concepts within hidden states. SafeSpeech [78] proactively affects voice data during training to make synthesized audio less usable, achieving robustness against voice cloning. DeBackdoor [79] 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. [80] document that extracting pre-trained encoders may cost substantially less than training from scratch. Lu et al. [81] 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.

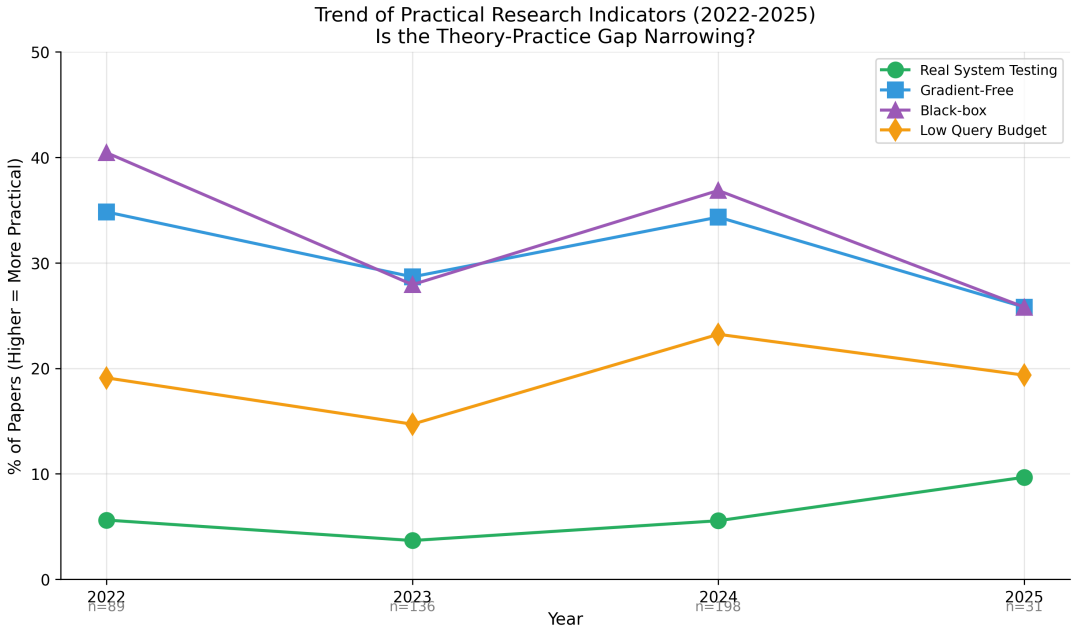

Figure 10: Temporal trends showing the theory-practice gap from 2022 to 2025, with limited change in key metrics.

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 [82]. 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

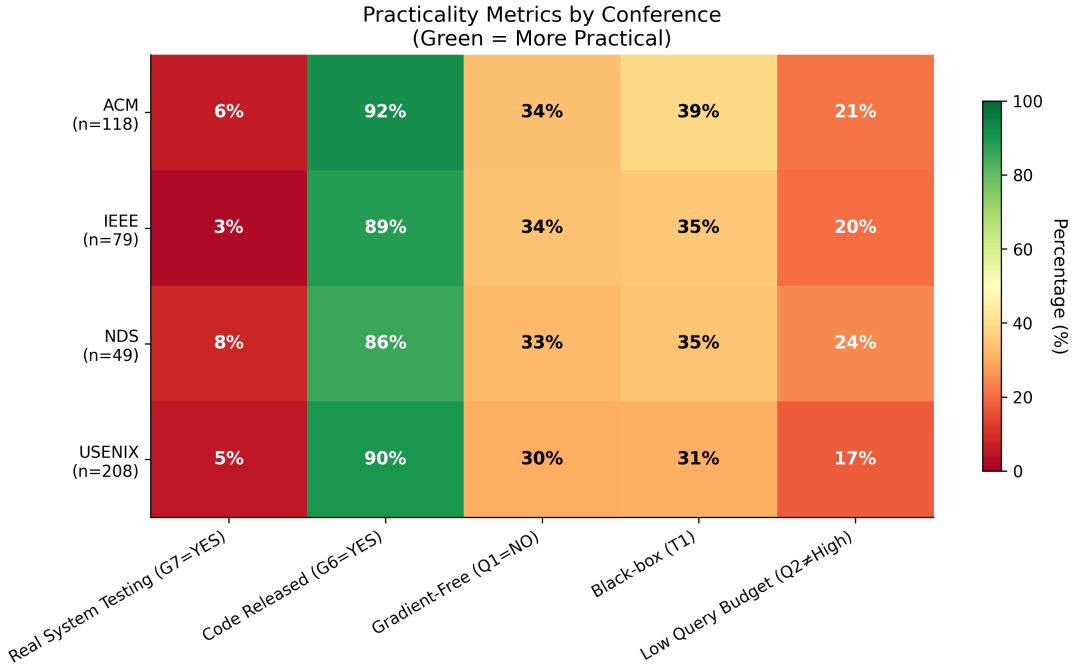

Figure 11: Conference comparison heatmap showing similar patterns across all venues.

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 [80, 81], documenting relationships between mathematical metrics and human perception [64], analyzing system-level integration considerations [71], 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 [40], documented accuracy-privacy trade-offs [83], and examined scalability considerations for certified robustness deployment [84].

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 [45], developing metrics for non-image domains [56], analyzing distributed training heterogeneity [85], 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 [57, 58], analyzed trade-offs involving computational cost, regulatory compliance, and usability [59], studied adversarial dynamics when defenses encounter adaptive attackers in production, and examined domain-specific requirements [50].

7.5 Attack versus Defense Papers

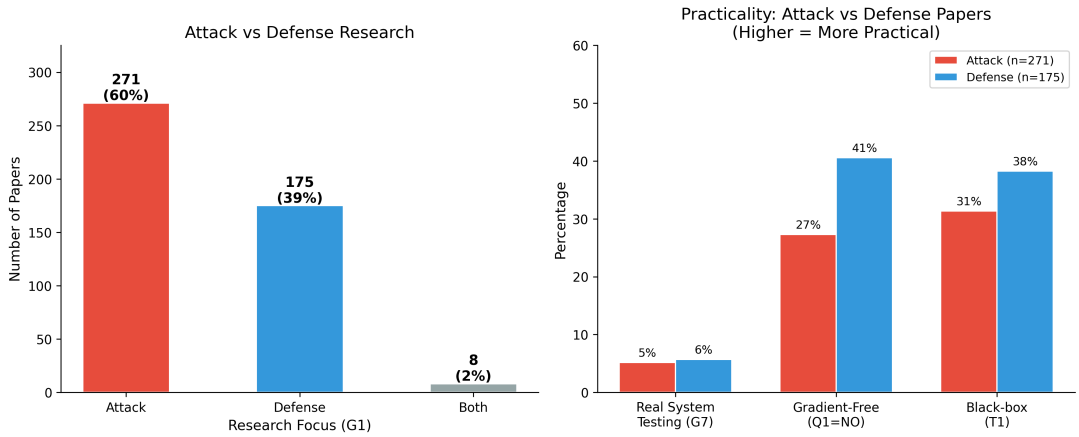

Figure 12: Comparison of practicality metrics between attack-focused and defense-focused papers.

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

Our analysis reveals a persistent theory-practice gap: 94.7% of papers avoid real-system testing, 67.8% require gradient access, and 63.2% assume white-box knowledge. Closing this gap requires coordinated intervention across the research ecosystem. We provide evidence-based recommendations for four stakeholder groups, grounded in recent frameworks and successful initiatives from the security and machine learning communities.

8.1 For Researchers

8.1.1 Specify Realistic Threat Models

The prevalence of white-box assumptions (63.2% of papers) contradicts deployment realities where models are protected behind APIs and security controls. Researchers need structured vocabularies to specify attacker capabilities precisely. The NIST AI Risk Management Framework provides such a taxonomy, classifying attacks by attacker knowledge level, lifecycle stage, and objectives [86]. Importantly, NIST notes that most documented attacks require minimal system knowledge, contradicting the white-box default in academic work.

MITRE ATLAS extends the ATT&CK framework to AI systems with 66 documented techniques and 33 real-world case studies [87]. These case studies reveal that successful attacks, including the 2024 Morris II worm [88] and the 2025 state-sponsored AI agent campaign [15], succeeded without gradient access or white-box knowledge. Researchers should map proposed attacks to ATLAS techniques and justify when assumptions exceed documented attacker capabilities. When white-box access is necessary for theoretical contributions, authors should explicitly discuss the gap between their assumptions and deployment constraints.

8.1.2 Employ Realistic Evaluation Frameworks

Evaluation must enforce constraints that deployed systems actually face. For adversarial robustness in computer vision, AutoAttack has become the standard because it revealed that 13 of 50 evaluated defenses had actual robustness at least 10% lower than initially reported [89]. The ensemble includes gradient-free attacks, providing a model for combining white-box and black-box evaluation. RobustBench maintains standardized leaderboards with over 120 models, requiring non-zero gradients to prevent gradient masking [90].

Query budgets present another disconnect. Among papers in our dataset that involve model queries, 80.4% assume budgets exceeding 1000 queries. Commercial APIs implement rate limiting, costs accumulate with usage, and repeated similar queries trigger anomaly detection. Query-efficient evaluation should become standard practice, with researchers reporting both the number of queries required and whether this budget reflects operational constraints.

For large language model security, standardized benchmarks are emerging. HarmBench provides 400 harmful behaviors across seven categories with explicit evaluation protocols [91], while JailbreakBench offers threat model specifications and scoring functions that enable reproducible comparison [92]. These benchmarks demonstrate how standardization can improve both rigor and practical relevance.

8.1.3 Report Practical Attack Economics

Single-attempt success rates misrepresent operational risk. Recent industry evaluations report success rates across multiple attempts (1, 50, 200), revealing dramatic increases with attacker persistence [93]. The 2024 SaTML Capture-the-Flag competition, involving over 137,000 adversarial interactions against 44 defenses, found that every defense was eventually bypassed through multi-turn attacks [94]. Single-turn evaluation would have missed this critical finding.

Researchers should report: (1) computational costs and query budgets required for attacks, (2) multi-turn success rates where applicable, (3) transferability across model families, and (4) economic costs to mount attacks. For domains beyond images, the constraints differ. Pierazzi et al. demonstrate how to formalize domain-specific constraints for malware, explicitly modeling semantic preservation and side effects [95]. Their work provides a template for extending adversarial ML beyond L_p perturbations.

8.2 For Conferences and Reviewers

8.2.1 Require Artifact Availability

While 89.6% of papers in our dataset release code, reproducibility requires more than source availability. USENIX Security now mandates that papers share artifacts on permanent repositories (Zenodo, FigShare, Software Heritage) or provide justification for non-release [96]. This policy establishes a baseline without excessive burden. All four surveyed venues offer three-tier badge systems (Available, Functional, Reproduced), but adoption varies. NDSS introduced artifact evaluation only in 2024, while USENIX and IEEE S&P have established programs.

We recommend that security venues require artifact availability as a publication condition, with graduated evaluation for functionality and reproducibility. Permanent archival with DOI assignment ensures long-term accessibility. The REFORMS checklist provides 32 questions covering problem specification, data handling, and limitations that can guide both authors and reviewers [97].

8.2.2 Evaluate Threat Model Realism

Review criteria should explicitly address assumptions underlying adversarial claims. We propose that authors specify: (1) attacker knowledge using NIST taxonomy categories, (2) query budgets and whether they reflect rate limiting and detection, (3) physical realizability for attacks on deployed systems, (4) economic viability including cost-benefit analysis, and (5) adaptive adversary considerations for defense papers.

The SaTML conference welcomes position papers addressing methodological concerns [98], creating space for research that advances evaluation methodology. Other venues should consider similar tracks. The AISec Workshop, co-located with CCS for 18 consecutive years, demonstrates sustained demand for work bridging ML and security [99]. Expanding industry review tracks with practitioner input can help identify when assumptions diverge from operational reality.

8.3 For Practitioners

8.3.1 Implement Defense in Depth

Academic papers typically evaluate single defenses in isolation. Deployed systems require layered protection because no single defense is perfect. Google's Secure AI Framework provides a structured approach with four pillars: secure development, deployment, execution, and monitoring [100]. Each pillar includes specific controls and risk assessments.

The OWASP Top 10 for LLM Applications establishes community consensus on priorities [101]. Prompt injection tops the list because, unlike many academic attacks, it requires no model access or technical sophistication. The 2025 update for agentic systems adds risks specific to multi-agent architectures, including memory poisoning and tool manipulation [102]. Practitioners should design systems assuming prompt injection will succeed, implementing architectural isolation that separates control logic from data processing [103].

8.3.2 Conduct Continuous Adversarial Testing

Point-in-time security assessments miss evolving threats. Microsoft's PyRIT automates red teaming with curated attack datasets, enabling continuous evaluation [104]. Open-source alternatives like Promptfoo and DeepEval provide similar capabilities with OWASP alignment [105, 106]. These tools can integrate into CI/CD pipelines, making adversarial testing part of standard development rather than a separate security audit.

Multi-turn attacks achieve over 90% success against defenses showing near-zero vulnerability in single-turn evaluation [107]. This finding underscores why continuous testing matters. Organizations should establish internal red teams that probe systems throughout the development lifecycle, not just before deployment.

8.3.3 Contribute to Shared Threat Intelligence

MITRE ATLAS contains 33 real-world case studies because practitioners documented incidents [87]. Organizations that experience novel attacks should consider contributing to ATLAS or the AI Incident Database. The Coalition for Secure AI, with members including Google, Microsoft, Amazon, Anthropic, and OpenAI, provides venues for pre-competitive collaboration on supply chain security and risk governance [108]. Threat intelligence sharing benefits defenders more than attackers because defenses require comprehensive coverage while attacks need only find one vulnerability.

8.4 For Funders and Policymakers

8.4.1 Require Deployment Validation

DARPA's GARD program required scenario-based evaluations connecting robustness claims to operational contexts [109]. The program produced the Adversarial Robustness Toolbox and APRICOT benchmark, both designed for practical application. The successor SABER program explicitly targets the "practical demonstration gap" by funding both research teams and operational assessment teams [110].

The NSF National AI Research Institutes program awards 16 to 20 million dollars over four to five years, sufficient for sustained work rather than isolated projects [111]. Theme 6 on AI and cybersecurity

requires integration of fundamental advances with practical security problems. This structure provides a model: funding should prioritize proposals that include deployment validation components, partnerships with operational organizations, and plans for transitioning research to practice.

8.4.2 Support Evaluation Infrastructure

Individual research groups cannot each build comprehensive evaluation platforms. Shared infrastructure reduces duplication and enables fair comparison. NIST's Dioptra testbed provides open-source infrastructure for evaluating AI security [112]. Its modular design allows researchers to swap datasets, models, attacks, and defenses. The UK AI Safety Institute's Systemic Safety Grants Programme funds infrastructure for evaluating deepfakes, misinformation, and system failures [113].

The SPHERE Research Infrastructure, used by NDSS for artifact evaluation, demonstrates how standardized testbeds enable reproducibility while reducing reviewer burden [114]. Funders should support such shared platforms rather than expecting each institution to develop parallel infrastructure. This investment pays dividends through improved reproducibility and reduced barriers to entry for new researchers.

8.4.3 Address Critical Gaps

Our analysis identifies three underfunded areas. First, human factors research represents only 0.09% of papers despite growing deployment in human-AI systems [115]. Attacks exploiting human cognition, timing, and trust may be more practical than mathematical perturbations. Second, economic analysis remains limited despite Apruzzese et al.'s argument that cost-driven threat modeling is essential [26]. Recent work provides foundations [116], but domain-specific models for fraud, malware, and other applications require development. Third, multi-turn and agentic evaluation lags behind deployment [117]. Systems combining multiple models, tools, and external services demand compositional security analysis that current benchmarks do not provide.

8.4.4 Align Regulation with Research

The EU AI Act, fully applicable in August 2026, requires high-risk systems to be "accurate, robust, and cybersecurity-resilient" with protection against adversarial attacks [118]. This creates demand for evaluation methodologies that translate academic benchmarks into compliance-ready protocols. The Act's regulatory sandboxes provide testing environments for this translation. Similarly, NIST's AI Risk Management Framework establishes governance structures that organizations use for risk assessment. The forthcoming Control Overlays for Securing AI Systems will adapt security controls specifically for AI vulnerabilities [119].

Regulation can drive practical research when compliance requires demonstrated robustness. Funders should support work that bridges academic benchmarks and regulatory requirements, ensuring that compliance obligations incentivize rather than hinder security research.

8.5 Path Forward

The theory-practice gap we document is not inevitable. Infrastructure for practical adversarial ML research exists through standardized frameworks (NIST AI RMF, MITRE ATLAS), evaluation tools (Au-

toAttack, RobustBench, HarmBench), shared testbeds (Dioptra, SPHERE), and industry collaboration venues (CoSAI). The challenge is adoption. Researchers can choose realistic threat models, conferences can require artifact availability and threat model justification, practitioners can implement defense in depth, and funders can prioritize deployment validation. These actions are concrete and achievable. Together, they can ensure that adversarial ML research addresses both theoretical rigor and operational security needs, building on the substantial foundation that existing work has established while directing future effort toward bridging the remaining gaps.

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] 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.

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

[33] 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.

[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] 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.

[36] 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.

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

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

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

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

[45] 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.

[46] 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.

[47] 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.

[48] 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.

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

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

[51] 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.

[52] 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.

[53] 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.

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

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

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

[63] 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.

[64] 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.

[65] 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.

[66] 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.

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

[68] 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.

[69] 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.

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

[71] 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.

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

[73] 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.

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

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

[76] 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.

[77] 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.

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

[79] 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.

[80] 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.

[81] 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.

[82] 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.

[83] 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.

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

[85] 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.

[86] Apostol Vassilev, Alina Oprea, Alie Fordyce, and Hyrum Anderson. Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. Technical Report NIST AI 100-2e2023, National Institute of Standards and Technology, 2023. URL https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf.

[87] MITRE Corporation. MITRE ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems, 2024. URL https://atlas.mitre.org. Accessed: 2025-01-12.

[88] Ben Nassi, Yisroel Mirsky, and Yuval Elovici. ComWorm: A Generative AI-Powered Worm. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, 2024.

[89] Francesco Croce and Matthias Hein. Reliable Evaluation of Adversarial Robustness with an Ensemble of Diverse Parameter-Free Attacks. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

[90] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. RobustBench: A Standardized Adversarial Robustness Benchmark. *Datasets and Benchmarks Track, NeurIPS*, 2021.

[91] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. In *arXiv preprint arXiv:2402.04249*, 2024.

[92] Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, J. Zico Kolter, Kai Bai, et al. JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models. In *arXiv preprint arXiv:2404.01318*, 2024.

[93] Anthropic. Claude Model Evaluations, 2024. URL https://www.anthropic.com/research. Technical report.

[94] Sander Schulhoff, Chenlu Wen, Yuvraj Parekh, Norman Mu, and Dan Hendrycks. The SaTML '24 CNN Classifier Backdoor and LLM Prompt Injection Competitions. In *SaTML 2024 Workshop*, 2024.

[95] Fabio Pierazzi, Feargus Pendlebury, Jacopo Cortellazzi, and Lorenzo Cavallaro. Intriguing Properties of Adversarial ML Attacks in the Problem Space. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, pages 1332–1349, 2020.

[96] USENIX. USENIX Security '25 Call for Artifacts, 2025. URL https://www.usenix.org/conference/usenixsecurity25/call-for-artifacts.

[97] 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. *Journal of Machine Learning Research*, 22(164):1–20, 2021.

[98] IEEE SaTML. IEEE Conference on Secure and Trustworthy Machine Learning (SaTML) 2025 Call for Papers, 2025. URL https://satml.org/participate-cfp/.

[99] Proceedings of the 17th ACM Workshop on Artificial Intelligence and Security. In *AISec '24*. ACM, 2024.

[100] Google Cloud. Secure AI Framework (SAIF). Technical report, 2024. URL https://saif.google/secure-ai-framework.

[101] OWASP Foundation. OWASP Top 10 for Large Language Model Applications v2.0, 2025. URL https://genai.owasp.org/.

[102] OWASP GenAI Security Project. OWASP Top 10 for Agentic AI Security, 2025. URL https://genai.owasp.org/2025/12/09/owasp-genai-security-project-releases-top-10-risks-and-mitigations-for-ager

[103] Richard Harang and Hervé Debar. Defending Against Prompt Injection Attacks Through Isolation and Defense in Depth. *arXiv preprint arXiv:2507.13169*, 2024.

[104] Microsoft Security Response Center. PyRIT: Python Risk Identification Toolkit for Generative AI, 2024. URL https://github.com/Azure/PyRIT.

[105] Promptfoo. Promptfoo: LLM Evaluation and Red Teaming Framework, 2024. URL https://www.promptfoo.dev/.

[106] Confident AI. DeepEval: The LLM Evaluation Framework, 2024. URL https://github.com/confident-ai/deepeval.

[107] Cem Anil, Daniel Deutsch, Ethan Perez, et al. Many-Shot Jailbreaking. Anthropic Technical Report, 2024.

[108] OASIS Open. Coalition for Secure AI (CoSAI), 2024. URL https://oasis-open.org/coalitions/secure-ai/.

[109] DARPA. DARPA GARD: Guaranteeing AI Robustness Against Deception, 2024. URL https://www.darpa.mil/research/programs/guaranteeing-ai-robustness-against-deception.

[110] DARPA. DARPA SABER: Securing Artificial Intelligence for Battlefield Effective Robustness, 2024. URL https://www.darpa.mil/research/programs/saber-securing-artificial-intelligence.

[111] National Science Foundation. NSF National Artificial Intelligence Research Institutes Program Solicitation, 2022.

[112] NIST. Dioptra: A Software Engineering Approach to Machine Learning Security Testbed, 2024. URL https://pages.nist.gov/dioptra/.

[113] UK AI Safety Institute. UK AI Safety Institute Systemic Safety Grants Programme, 2024. URL https://www.gov.uk/government/news/research-programme-to-ensure-uk-economy-uses-ai-to-grow-safely.

[114] SPHERE Consortium. SPHERE: Enabling Reproducibility through Research Infrastructure, 2024. URL https://www.ndss-symposium.org/wp-content/uploads/2025-poster-31.pdf.

[115] Jinghan Yu, Yang Yang, Xiaozhuan Huang, et al. Position: Human Factors Reshape Adversarial Analysis in Human-AI Decision-Making Systems. *arXiv preprint arXiv:2509.21436*, 2025.

[116] Kathrin Grosse, Sotirios Terzis, Tom Chothia, and Matthew Smith. On the Economics of Adversarial Machine Learning. *IEEE Transactions on Dependable and Secure Computing*, 2024. doi: 10.1109/TDSC.2024.3376499.

[117] Daizong Xu and Keshab K. Parhi. A Survey of Attacks on Large Vision-Language Models: Resources, Advances, and Future Trends. *arXiv preprint arXiv:2407.07403*, 2025.

[118] European Parliament and Council. Regulation (EU) 2024/1689 of the European Parliament and of the Council on Artificial Intelligence (AI Act), 2024. URL https://artificialintelligenceact.eu/.

[119] NIST. NIST COSAIS: Control Overlays for Securing AI Systems, 2024. Public draft forthcoming FY2026.

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	ExamplesforLeaderd	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	Illinois	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	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

Continued on next page

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		×	H	L	0	2

Continued on next page

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

Continued on next page

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	F	×	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

Continued on next page

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	Tradition	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	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	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

Continued on next page

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