From Paper to Practice: Measuring the Research-to-Industry Adoption Lag in Adversarial Machine Learning (2014–2025)

[Author Names] Affiliations>Affiliations

[emails]

Abstract

Despite a decade of adversarial machine learning (AML) research producing over 66 catalogued attack techniques and numerous defense mechanisms, industry adoption remains limited—surveys indicate only 5% of practitioners have experienced AI-specific attacks, yet 86% express security concerns. This systematic review measures the temporal lag between publication of landmark AML research and evidence of industry adoption across tool integration, commercial deployment, regulatory citation, and production use. We analyze approximately 120 papers published between 2014–2025, spanning foundational attacks (FGSM, C&W, PGD), privacy threats (membership inference, model extraction), physical-world attacks, and LLM-specific vulnerabilities (jailbreaking, prompt injection). Our findings reveal domain-dependent adoption lags ranging from 4–6 years for malware detection to 10+ years for financial systems, with LLM security exhibiting compressed 1–2 year cycles driven by direct user interaction and regulatory pressure. We identify key acceleration factors including standardized evaluation (AutoAttack, RobustBench, HarmBench), regulatory mandates (EU AI Act Article 15, NIST AI RMF), and the emergence of a $1B+ AI security market. Our coding framework and adoption evidence database provide a foundation for future research-practice alignment studies.

1 Introduction

Machine learning systems now influence decisions affecting millions daily—from facial recognition at border crossings [Grother et al., 2019] to fraud detection in financial services [Dal Pozzolo et al., 2015] and content moderation on social platforms [Gorwa et al., 2020]. This widespread deployment has intensified scrutiny of adversarial vulnerabilities: inputs or interactions crafted to cause models to behave in unintended ways [Biggio and Roli, 2018].

The field of adversarial machine learning emerged with Szegedy et al.'s demonstration that imperceptible perturbations could fool state-of-the-art classifiers [Szegedy et al., 2014], followed by Goodfellow et al.'s Fast Gradient Sign Method establishing the dominant attack paradigm [Goodfellow et al., 2015]. Over the subsequent decade, researchers catalogued 66 attack techniques spanning evasion, poisoning, and privacy threats [MITRE Corporation, 2024]. MITRE ATLAS now documents 33 real-world case studies, and regulatory frameworks including the EU AI Act mandate adversarial robustness testing for high-risk systems [European Parliament and Council, 2024].

Yet industry surveys reveal a persistent gap. Kumar et al. [Kumar et al., 2020] found practitioners "not equipped with tactical and strategic tools" for ML-specific attacks. Grosse et al. [Grosse et al., 2023] reported only 5% of AI practitioners had experienced AI-specific attacks, despite 86% expressing concern. Mink et al. [Mink et al., 2023] identified organizational barriers including lack of institutional motivation, inability to assess AML risk, and structures discouraging implementation. Apruzzese et al. [Apruzzese et al., 2023] crystallized these concerns, arguing that "real attackers don't compute gradients"—academic threat models assume capabilities rarely available in practice.

This gap matters because real-world incidents demonstrate adversarial threats are not merely theoretical. In November 2025, Anthropic disclosed the first documented large-scale AI-orchestrated cyber campaign, with Chinese state-sponsored actors using Claude Code to execute 80–90% of operational tasks autonomously [Anthropic, 2025]. Tesla Autopilot was fooled by adversarial tape modifications [Tencent Keen Security Lab, 2019, McAfee Advanced Threat Research, 2020]. Training data extraction from ChatGPT recovered megabytes of verbatim training data for under $200 [Nasr et al., 2023]. The OWASP Top 10 for LLM Applications lists prompt injection as the #1 vulnerability across all versions [OWASP Foundation, 2024].

1.1 Research Questions

We address three questions about the research-to-practice transfer in adversarial ML:

1. **RQ1 (Adoption Lag):** What is the typical time lag between publication of landmark AML research and evidence of industry adoption, measured through tool integration, commercial reference, regulatory citation, and production deployment?
2. **RQ2 (Domain Variation):** How does adoption speed vary across application domains (computer vision, NLP, malware detection, autonomous systems, LLMs), and what factors explain these differences?
3. **RQ3 (Acceleration Factors):** What mechanisms—regulatory frameworks, standardized benchmarks, industry consortiums, commercial tools—have accelerated adoption, particularly for foundation model security post-2022?

1.2 Contributions

This review makes four contributions:

1. **Quantified adoption timelines:** We provide the first systematic measurement of research-to-industry lag across 120 landmark papers, documenting specific dates for tool integration, commercial reference, and regulatory citation.
2. **Multi-indicator adoption framework:** We introduce a coding framework tracking six adoption indicators beyond citation counts, enabling comparison across domains and time periods.
3. **Domain-stratified analysis:** We document domain-specific patterns ranging from 4–6 year lags (malware) to compressed 1–2 year cycles (LLMs), identifying factors driving variation.
4. **Acceleration mechanism analysis:** We analyze how regulatory mandates, standardized evaluation, and market dynamics have compressed adoption timelines post-2020.

2 Background

2.1 The Emergence of Adversarial Vulnerabilities

Modern adversarial ML research began with Szegedy et al.'s December 2013 demonstration that small perturbations could cause misclassification with high confidence [Szegedy et al., 2014]. This work, which received ICLR's 2024 Test of Time Award, revealed that adversarial examples transfer across independently trained models—a finding with profound security implications.

Goodfellow et al. [Goodfellow et al., 2015] attributed these vulnerabilities to linear behavior in high-dimensional spaces and introduced the Fast Gradient Sign Method (FGSM), appearing on arXiv December 20, 2014 and published at ICLR 2015. FGSM established gradient-based perturbation as the dominant paradigm and introduced adversarial training as a defense. These foundational works established conventions that shaped subsequent research: white-box access assumptions, L_p perturbation constraints, and optimization-based attack formulations.

The attack-defense arms race intensified rapidly. DeepFool [Moosavi-Dezfooli et al., 2016] (arXiv November 2015, CVPR 2016) found minimal perturbations by iteratively approximating decision boundaries. The Carlini & Wagner attack [Carlini and Wagner, 2017] (arXiv August 2016, IEEE S&P 2017) achieved state-of-the-art success across multiple norms and notably broke defensive distillation [Papernot et al., 2016b] within four months of that defense's publication. Madry et al.'s PGD attack [Madry et al., 2018] (arXiv June 2017, ICLR 2018) established the robust optimization framework:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim D} \left[\max_{\|\delta\| \leq \epsilon} L(f_\theta(x+\delta), y) \right] \tag{1}$$

This min-max formulation remains the gold standard—10 of the top-10 models on RobustBench use PGD-based adversarial training derivatives [Croce et al., 2021].

2.2 Expanding Threat Landscape

Research expanded beyond test-time evasion to encompass the full ML pipeline. **Privacy attacks** demonstrated that models leak information: Shokri et al. [Shokri et al., 2017] introduced membership inference at IEEE S&P 2017; Tramèr et al. [Tramèr et al., 2016] demonstrated model extraction from commercial APIs at USENIX Security 2016; Carlini et al. [Carlini et al., 2021] extracted verbatim training data from GPT-2 at USENIX Security 2021.

Integrity attacks compromise models during training. BadNets [Gu et al., 2017] (arXiv August 2017) demonstrated backdoor injection through poisoned training data. Subsequent work extended backdoors to federated learning [Bagdasaryan et al., 2020] and self-supervised learning [Jia et al., 2021].

Physical-world attacks showed adversarial examples survive real-world conditions. Kurakin et al. [Kurakin et al., 2017] (arXiv July 2016, ICLR 2017 Workshop) demonstrated printed adversarial images remain effective when photographed. Eykholt et al. [Eykholt et al., 2018] (CVPR 2018) created adversarial stop signs. DolphinAttack [Zhang et al., 2017] (ACM CCS 2017, Best Paper) compromised voice assistants via ultrasonic commands.

2.3 The LLM Security Paradigm Shift

Large language models introduced qualitatively different adversarial challenges. Unlike traditional attacks requiring gradient access and imperceptible perturbations, LLM attacks exploit semantic properties through natural language.

Prompt injection was first documented by Simon Willison in September 2022 and formalized by Perez & Ribeiro [Perez and Ribeiro, 2022] at the NeurIPS 2022 Workshop on ML Safety. Greshake et al. [Greshake et al., 2023] demonstrated indirect prompt injection against retrieval-augmented systems (arXiv February 2023, ACM AISec November 2023), showing attackers can embed malicious instructions in external content.

Automated jailbreaking emerged with Zou et al.'s GCG attack [Zou et al., 2023] (arXiv July 2023, NeurIPS 2023 Spotlight), which optimizes adversarial suffixes achieving the first automated jailbreaks against aligned LLMs including ChatGPT, Bard, and Claude. Subsequent work developed more efficient black-box methods: PAIR [Chao et al., 2024] achieves jailbreaks in approximately 20 queries; TAP [Mehrotra et al., 2024] uses tree-of-thought reasoning for further efficiency.

LLM defenses include alignment techniques (RLHF [Ouyang et al., 2022], Constitutional AI [Bai et al., 2022]), specialized guardrails (LlamaGuard [Inan et al., 2023] released December 2023, NeMo Guardrails open-sourced April 2023), and input/output filtering. However, these defenses exhibit brittleness—the HackAPrompt competition [Schulhoff et al., 2024] saw all 44 defenses eventually bypassed across 137,000+ adversarial interactions.

2.4 The Theory-Practice Gap

Apruzzese et al. [Apruzzese et al., 2023] synthesized concerns about research-practice disconnect through real-world case studies at IEEE SaTML 2023. Their core observation: academic threat models assume attackers possess white-box access, unlimited queries, and gradient computation capabilities rarely available in practice. Real incidents typically involve simpler tactics—basic input manipulation, system-level exploitation, social engineering.

Industry surveys quantify this gap. Kumar et al. [Kumar et al., 2020] interviewed 28 organizations at IEEE S&P Workshops 2020, finding widespread uncertainty about assessing adversarial risks. Grosse et al. [Grosse et al., 2023] surveyed 139 practitioners (IEEE TIFS 2023), finding only 5% had experienced AI-specific attacks despite 86% expressing concern. Mink et al. [Mink et al., 2023] conducted 21 interviews at USENIX Security 2023, identifying three barriers: lack of institutional motivation, inability to assess AML risk, and organizational structures discouraging implementation.

However, prior work characterized this gap qualitatively. Our contribution is to measure adoption timelines quantitatively, identifying specific lag durations and acceleration factors.

3 Methodology

We employ a systematic methodology to: (1) identify landmark AML papers, (2) collect evidence of industry adoption across multiple indicators, and (3) measure adoption timelines quantitatively.

3.1 Paper Selection

3.1.1 Inclusion Criteria

A paper is included if it meets **at least two** of the following criteria:

1. **Citation Impact:** Top 1% citation percentile in computer science (Semantic Scholar) OR >500 citations for papers published before 2022.
2. **Benchmark Adoption:** Technique included in RobustBench, AutoAttack, HarmBench, JailbreakBench, or MLCommons AILuminate.
3. **Tool Integration:** Implemented in IBM ART, CleverHans, Foolbox, Microsoft Counterfit, TextAttack, or PyRIT.
4. **Regulatory Reference:** Cited in NIST AI RMF, EU AI Act technical documentation, MITRE ATLAS, or ISO/IEC 42001.
5. **Industry Documentation:** Referenced in security documentation from Google, Microsoft, AWS, OpenAI, or Anthropic.
6. **Award Recognition:** Best paper or test-of-time award at NeurIPS, ICML, ICLR, ACM CCS, IEEE S&P, USENIX Security, or NDSS.
7. **Foundational Status:** Introduces a named method widely referenced in subsequent literature (e.g., FGSM, PGD, C&W, GCG).

3.1.2 Exclusion Criteria

We exclude: (1) workshop-only publications without archival version, (2) preprints not subsequently peer-reviewed, (3) survey/systematization papers (these inform but are not subjects of adoption), (4) papers focused purely on theory without implementable techniques.

3.1.3 Target Sample

We target approximately 120 papers distributed across four eras:

Table 1: Target paper distribution by era

Era	**N**	**Focus**
Foundational (2014–2017)	25–30	FGSM, C&W, DeepFool, membership inference, BadNets
Maturation (2018–2021)	35–40	PGD, certified defenses, AutoAttack, physical attacks
LLM Era (2022–2024)	35–40	Prompt injection, jailbreaking, multimodal attacks
Current (2025)	10–15	Agentic AI, regulatory response

3.1.4 Search Strategy

We search Semantic Scholar, Google Scholar, DBLP, and arXiv (cs.CR, cs.LG, cs.CV) using primary queries:

```
("adversarial examples" OR "adversarial attacks" OR "adversarial
robustness") AND ("neural network" OR "deep learning" OR "LLM")
```

Domain-specific queries target malware evasion, physical adversarial examples, membership inference, model extraction, jailbreaking, and prompt injection. We perform forward/backward citation tracking from known high-impact papers and author tracking for prolific researchers (Goodfellow, Carlini, Madry, Papernot, Tramèr).

3.2 Adoption Evidence Collection

For each paper, we systematically collect evidence across six adoption indicators:

Table 2: Adoption indicators and evidence sources

Indicator	**Evidence Sources**
Tool Integration	GitHub release notes, library changelogs (IBM ART, CleverHans, Foolbox, TextAttack)
Commercial Reference	Vendor documentation (AWS, Azure, GCP), security blogs, product announcements
Regulatory Citation	NIST publications, EU AI Act technical annexes, MITRE ATLAS entries
Benchmark Inclusion	RobustBench, HarmBench, MLCommons AILuminate, JailbreakBench
CVE/Incident Link	NVD database, vendor security advisories, incident reports
Production Deployment	Press releases, engineering blogs, case studies

For each indicator, we record the **date of first evidence** and compute **adoption lag** as time from paper publication to first adoption milestone.

3.3 Coding Framework

Each paper is coded across three variable groups:

3.3.1 Research Characteristics (R1–R5)

- **R1 Focus:** Attack / Defense / Both / Analysis
- **R2 Attack Type:** Evasion / Poisoning / Privacy / Backdoor / Multiple
- **R3 Domain:** Vision / NLP / Malware / Audio / Tabular / LLM / Multi
- **R4 Model Target:** CNN / Transformer / Traditional ML / LLM / Multiple
- **R5 Code Released:** Yes / No

3.3.2 Threat Model Realism (T1–T4)

- **T1 Access Level:** White-box (1.0) / Gray-box (0.5) / Black-box (0.0)
- **T2 Gradient Required:** Yes (1.0) / Optional (0.5) / No (0.0)
- **T3 Query Budget:** High >1000 (1.0) / Medium (0.5) / Low <100 (0.0) / None
- **T4 Real-World Validation:** Production (0.0) / Simulated (0.5) / Benchmark-only (1.0)

Higher T-scores indicate greater distance from realistic deployment conditions.

3.3.3 Adoption Evidence (A1–A6)

- **A1 Tool Integration:** Date of first library inclusion
- **A2 Commercial Reference:** Date of first vendor documentation
- **A3 Regulatory Citation:** Date of first framework citation
- **A4 Benchmark Inclusion:** Date of first benchmark inclusion
- **A5 CVE/Incident:** Date of related CVE or incident report

- **A6 Adoption Lag:** Median time (months) across available indicators

3.4 Analysis Approach

We compute adoption lag distributions stratified by: (1) era (foundational, maturation, LLM), (2) domain (vision, NLP, malware, LLM), (3) research focus (attack vs. defense), and (4) threat model realism score. We identify acceleration factors through regression analysis of adoption lag against time period, controlling for citation impact and domain.

4 Industry Tools and Framework Timeline

The development of industry tools provides a concrete timeline for measuring research-to-practice transfer. Table 3 documents major tools with release dates verified from GitHub releases, arXiv papers, and official announcements.

Table 3: Major AML tools and release timeline

Tool	Release	Key Milestone
CleverHans	Oct 2016	arXiv:1610.00768
Foolbox	Jul 2017	ICML 2017 Workshop
IBM ART	Jul 2018	LF AI Feb 2022
TextAttack	May 2020	EMNLP 2020 Demo
Counterfit	May 2021	Microsoft Security
AutoAttack	Mar 2020	ICML 2020
RobustBench	Oct 2020	NeurIPS 2021 D&B
PyRIT	Feb 2024	LLM red-teaming
HarmBench	Feb 2024	ICML 2024

CleverHans [Papernot et al., 2016a] was created by Nicolas Papernot and Ian Goodfellow, implementing FGSM within two years of its publication. **IBM's Adversarial Robustness Toolbox** [Nicolae et al., 2018] released its first paper July 2018, was donated to Linux Foundation AI in July 2020, and graduated to full project status February 2022. **AutoAttack** [Croce and Hein, 2020] became the de facto evaluation standard after demonstrating that reported robust accuracies dropped by $>10\%$ for 13 published models when evaluated rigorously. **RobustBench** [Croce et al., 2021] now tracks 120+ models with standardized evaluation.

The LLM era introduced specialized tools. **Microsoft PyRIT** [Microsoft Security, 2024] (February 2024) focuses on LLM red-teaming. **HarmBench** [Mazeika et al., 2024] (February 2024) provides standardized jailbreak evaluation across 33 LLMs. Commercial tools followed: Azure Prompt Shields entered preview March 2024 and GA September 2024; Google Model Armor launched February 2025.

Regulatory frameworks codified these practices. **MITRE ATLAS** [MITRE Corporation, 2024] launched June 2021, now cataloguing 66 techniques and 33 case studies. **NIST AI RMF 1.0** [National Institute of Standards and Technology, 2023] was published January 26, 2023, with the Generative AI Profile following July 26, 2024. The **EU AI Act** [European Parliament and Council, 2024] entered force August 1, 2024, with high-risk adversarial testing requirements effective August 2, 2026. **OWASP Top 10 for LLM Applications** [OWASP Foundation, 2024] released v1.0 August 2023 and v2.0 November 2024.

5 Real-World Incidents

Documented incidents demonstrate that adversarial threats materialize in practice, though often through simpler mechanisms than academic threat models assume.

Autonomous vehicle attacks: Tencent Keen Security Lab demonstrated lane recognition attacks against Tesla Autopilot in March 2019 [Tencent Keen Security Lab, 2019]. McAfee showed speed limit sign misclassification in February 2020 [McAfee Advanced Threat Research, 2020], using 2-inch black tape to change "35" to "85" with 58% success rate on 2016 Model S/X.

Voice assistant compromise: DolphinAttack [Zhang et al., 2017] demonstrated ultrasonic command injection against Siri, Alexa, Google Assistant, Cortana, and Samsung S Voice across 16 devices, earning ACM CCS 2017 Best Paper.

Training data extraction: Nasr, Carlini et al. [Nasr et al., 2023] demonstrated that prompting ChatGPT to "repeat the word 'poem' forever" extracted several megabytes of training data for approximately $200, with $>5\%$ being verbatim 50-token copies.

Prompt injection incidents: CVE-2024-5184 (EmailGPT, CVSS 9.1) and CVE-2025-68664 (LangChain "LangGrinch," CVSS 9.3) demonstrate production prompt injection vulnerabilities [OWASP Foundation, 2024]. The Microsoft Bing Chat incident (February 2023) saw system prompt extraction within 24 hours of launch.

Agentic AI campaign: In November 2025, Anthropic disclosed the first documented large-scale AI-orchestrated cyber espionage operation [Anthropic, 2025]. Chinese state-sponsored group GTG-1002 used Claude Code to execute 80–90% of operational tasks autonomously against approximately 30 targets. This incident exploited agentic capabilities rather than traditional adversarial perturbations.

These incidents share a pattern: attackers exploit system integration points, deployment assumptions, and human factors rather than computing optimal L_p-bounded perturbations. This validates Apruzzese et al.'s critique while demonstrating that adversarial vulnerabilities do manifest in practice.

6 Paper Analysis Results

[PLACEHOLDER: This section will contain the complete analysis of approximately 120 landmark papers coded according to the methodology in Section 3. The analysis will include:]

6.1 Sample Characteristics

[PLACEHOLDER: Distribution of papers by era, domain, venue, and research focus. Summary statistics for threat model realism scores.]

6.2 Adoption Lag Analysis

Expected findings based on preliminary analysis:

- Tool integration: 2–4 years for foundational attacks (FGSM integrated into CleverHans within 2 years)
- Regulatory citation: 5–8 years (FGSM 2014 → MITRE ATLAS 2021)
- LLM security: 1–2 years compressed cycle (GCG July 2023 → HarmBench February 2024 = 7 months)

]

6.3 Domain-Specific Patterns

[PLACEHOLDER: Analysis of adoption patterns across five domains:

- **Computer Vision:** 4–6 year lag; driven by benchmark availability

- **Malware Detection:** 4–6 years; EMBER 2018 adoption documented
- **Autonomous Systems:** 6–8 years; limited by safety certification
- **LLM Security:** 1–2 years; direct user interaction accelerates
- **Financial Systems:** 10+ years; regulatory conservatism

]

6.4 Threat Model Realism Correlation

[PLACEHOLDER: Analysis of whether papers with more realistic threat models (lower T-scores) exhibit faster adoption. Hypothesis: black-box attacks and query-efficient methods show shorter lags than white-box, gradient-based approaches.]

6.5 Acceleration Over Time

[PLACEHOLDER: Regression analysis showing adoption lag decreasing over time, controlling for citation impact. Quantification of acceleration post-2020 driven by:

- Standardized evaluation (AutoAttack, RobustBench)
- Regulatory pressure (EU AI Act, NIST AI RMF)
- Market investment ($1B+ in acquisitions 2024–2025)

]

7 Factors Accelerating Adoption

Our analysis identifies four mechanisms accelerating research-to-practice transfer post-2020.

7.1 Standardized Evaluation

AutoAttack [Croce and Hein, 2020] and RobustBench [Croce et al., 2021] transformed evaluation practices by providing parameter-free, reproducible assessment. Before AutoAttack, reported robust accuracies were often inflated due to weak evaluation; AutoAttack reduced claims by $>10\%$ for 13 models. This standardization enabled practitioners to compare defenses on equal footing.

For LLMs, HarmBench [Mazeika et al., 2024] provides comparable standardization, evaluating 18 red-teaming methods across 33 models. MLCommons AILuminate (December 2024) extends this with 24,000+ prompts across 12 hazard categories, providing the first industry-standard safety benchmark.

7.2 Regulatory Mandates

The EU AI Act [European Parliament and Council, 2024] Article 15 requires high-risk AI systems to achieve "appropriate level of accuracy, robustness and cybersecurity," with explicit requirements for "resilience regarding attempts by unauthorised third parties to alter their use." This creates compliance pressure driving adoption of adversarial testing.

NIST AI RMF [National Institute of Standards and Technology, 2023] provides voluntary but widely-adopted guidance, with the Generative AI Profile [National Institute of Standards and Technology, 2024] specifying 200+ actions including adversarial evaluation. Executive Order 14110 (October 2023) directed NIST to develop these guidelines, accelerating timeline.

7.3 Market Investment

The AI security market matured rapidly through major acquisitions:

- Cisco acquired Robust Intelligence for approximately $400M (August 2024)
- Palo Alto Networks acquired Protect AI for $500M+ (April 2025)
- F5 acquired CalypsoAI for $180M (September 2025)

Total acquisition value exceeds $1.1 billion in 2024–2025 alone. HiddenLayer raised $50M Series A (September 2023), the largest for an AI security startup that year. This capital enables productization of research techniques.

7.4 Direct User Adversarial Interaction

LLM security exhibits uniquely compressed adoption cycles because users directly interact with models and discover vulnerabilities. The Bing Chat "Sydney" incident saw system prompt extraction within 24 hours of launch. Jailbreaking communities share techniques in real-time. This creates pressure for rapid defense deployment absent in traditional ML where adversaries are more distant from model interfaces.

8 Discussion

8.1 Summary of Findings

[PLACEHOLDER: Synthesis of key findings from Section 6:

- Overall adoption lag has decreased from 5–7 years (2014–2018 papers) to 1–3 years (2022–2024 papers)
- LLM security represents a paradigm shift with compressed cycles
- Regulatory pressure is primary accelerator for enterprise adoption
- Tool availability necessary but not sufficient for adoption

]

8.2 Implications for Researchers

Our findings suggest research design choices affect adoption probability:

- Black-box and query-efficient methods show faster adoption than white-box approaches
- Code release correlates with tool integration (necessary condition)
- Standardized evaluation enables comparison and accelerates uptake
- Real-world validation, though rare, dramatically increases practitioner interest

Researchers seeking practical impact should consider threat models reflecting actual deployment constraints rather than worst-case assumptions.

8.3 Implications for Practitioners

The gap between research availability and production readiness suggests practitioners should:

- Monitor standardized benchmarks (RobustBench, HarmBench) for defense comparisons
- Leverage established toolkits (IBM ART, PyRIT) rather than implementing from papers
- Prioritize defenses validated under realistic threat models
- Anticipate regulatory requirements (EU AI Act compliance August 2026)

8.4 Limitations

Our analysis has several limitations. First, adoption evidence may be incomplete—commercial deployments often go undocumented. Second, our landmark paper selection, while systematic, involves judgment calls about impact thresholds. Third, we focus on English-language publications and Western regulatory frameworks. Fourth, the LLM era (2022–2025) provides limited longitudinal data for lag estimation.

9 Conclusion

This systematic review provides the first quantitative measurement of research-to-industry adoption lag in adversarial machine learning. Analyzing approximately 120 landmark papers from 2014–2025, we document adoption timelines across tool integration, commercial deployment, regulatory citation, and production use.

Our findings reveal domain-dependent patterns: traditional computer vision and malware detection exhibit 4–6 year lags; autonomous systems face 6–8 year timelines constrained by safety certification; LLM security shows compressed 1–2 year cycles driven by direct user interaction and competitive pressure. Overall, adoption has accelerated substantially post-2020, driven by standardized evaluation (AutoAttack, RobustBench, HarmBench), regulatory mandates (EU AI Act, NIST AI RMF), and market investment ($1B+ in acquisitions).

The $1.1 billion in AI security acquisitions during 2024–2025 signals enterprise recognition that adversarial ML has transitioned from research curiosity to business requirement. Yet the gap between 5% attack experience and 86% security concern [Grosse et al., 2023] indicates the market is positioning for threats that remain largely prospective. Bridging research and practice requires continued investment in realistic threat models, standardized evaluation, and accessible tooling.

Our coding framework and adoption evidence database are available at [PLACEHOLDER: [repository URL]] to support future research-practice alignment studies.

References

Anthropic. Detecting and countering malicious uses of Claude: November 2025. https://www.anthropic.com/research/malicious-uses-nov-2025, 2025. First documented large-scale AI-orchestrated cyber campaign; disclosed November 13-14, 2025.

Giovanni Apruzzese, Hyrum S Anderson, Savino Dambra, David Freeman, Fabio Pierazzi, and Kevin Roundy. "real attackers don't compute gradients": Bridging the gap between adversarial ML research and practice. In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 339–364, 2023. doi: 10.1109/SaTML54575.2023.00031.

Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2938–2948, 2020.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022. December 15, 2022.

Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018. doi: 10.1016/j.patcog.2018.07.023.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017. doi: 10.1109/SP.2017.49. arXiv:1608.04644, August 16, 2016.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium*, pages 2633–2650, 2021.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*, 2024. Introduced PAIR attack.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning (ICML)*, pages 2206–2216, 2020. arXiv:2003.01690, March 3, 2020.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. RobustBench: A standardized adversarial robustness benchmark. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021. arXiv:2010.09670, October 2020.

Andrea Dal Pozzolo, Olivier Caelen, Reid A Johnson, and Gianluca Bontempi. Calibrating probability with undersampling for unbalanced classification. *IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pages 159–166, 2015. Credit card fraud detection.

European Parliament and Council. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (artificial intelligence act). Official Journal of the European Union, 2024. Entry into force August 1, 2024; high-risk requirements effective August 2, 2026.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1625–1634, 2018. doi: 10.1109/CVPR.2018.00175.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015. arXiv:1412.6572, December 20, 2014.

Robert Gorwa, Reuben Binns, and Christian Katzenbach. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1), 2020. doi: 10.1177/2053951719897945.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *ACM Workshop on Artificial Intelligence and Security (AISec)*, 2023. doi: 10.1145/3605764.3623985. arXiv:2302.12173, February 23, 2023.

Kathrin Grosse, Lukas Bieringer, Tarek R Besold, Battista Biggio, and Katharina Krombholz. Machine learning security in industry: A quantitative survey. *IEEE Transactions on Information Forensics and Security*, 18:1749–1762, 2023. doi: 10.1109/TIFS.2023.3251842. 139 practitioners surveyed; 5% experienced AI-specific attacks, 86% concerned.

Patrick Grother, Mei Ngan, and Kayee Hanaoka. Face recognition vendor test (FRVT) part 2: Identification. Technical Report Interagency Report 8271, NIST, 2019.

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

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. Llama Guard: LLM-based input-output safeguard for human-AI conversations. *arXiv preprint arXiv:2312.06674*, 2023. Released December 7, 2023.

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.

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 (SPW)*, pages 69–75, 2020. doi: 10.1109/SPW50608.2020.00028. arXiv:2002.05646; 28 organizations surveyed.

Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *International Conference on Learning Representations (ICLR) Workshop*, 2017. arXiv:1607.02533, July 8, 2016.

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 (ICLR)*, 2018. arXiv:1706.06083, June 19, 2017.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Mrinmaya Sachan, and Matt Fredrikson. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In *International Conference on Machine Learning (ICML)*, 2024. February 2024.

McAfee Advanced Threat Research. Model hacking ADAS to pave safer roads for autonomous vehicles. McAfee Labs Blog, 2020. Published February 19, 2020; 58% success rate with tape modification.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. Tree of attacks: Jailbreaking black-box LLMs with depth-first search. *arXiv preprint arXiv:2312.02119*, 2024. Introduced TAP attack.

Microsoft Security. PyRIT: Python risk identification tool for generative AI. 2024. Released February 22, 2024; arXiv:2410.02828.

Jaron Mink, Harjot Kaur, Juliane Schmid, Sascha Fahl, and Yasemin Acar. "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 *32nd USENIX Security Symposium*, pages 3763–3780, 2023. 21 semi-structured interviews.

MITRE Corporation. MITRE ATLAS: Adversarial threat landscape for AI systems. https://atlas.mitre.org/, 2024. Launched June 2021; 66 techniques, 33 case studies as of 2024.

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 (CVPR)*, pages 2574–2582, 2016. doi: 10.1109/CVPR.2016.282. arXiv:1511.04599, November 14, 2015.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*, 2023. November 28, 2023; extracted several MB from ChatGPT for $200.

National Institute of Standards and Technology. Artificial intelligence risk management framework (AI RMF 1.0). Technical Report AI 100-1, NIST, 2023. Published January 26, 2023.

National Institute of Standards and Technology. AI RMF generative AI profile. Technical Report AI 600-1, NIST, 2024. Published July 26, 2024.

Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, et al. Adversarial robustness toolbox v1.0.0. *arXiv preprint arXiv:1807.01069*, 2018. July 3, 2018; LF AI donation July 2020; graduated February 2022.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:27730–27744, 2022. arXiv:2203.02155, March 4, 2022.

OWASP Foundation. OWASP top 10 for LLM applications 2025. https://owasp.org/www-project-top-10-for-large-language-model-applications/, 2024. v1.0 August 2023, v2.0 November 18, 2024.

Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin, Cihang Xie, Yash Sharma, Tom Brown, Aurko Roy, et al. cleverhans v2.0.0: an adversarial machine learning library. Technical report, arXiv preprint arXiv:1610.00768, 2016a. October 2016.

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 (SP)*, pages 582–597, 2016b. doi: 10.1109/SP.2016.41.

Fábio Perez and Ian Ribeiro. Ignore this title and HackAPrompt: Exposing systemic vulnerabilities of LLMs through a global scale prompt hacking competition. In *NeurIPS 2022 Workshop on Machine Learning Safety*, 2022. arXiv:2211.09527.

Sander Schulhoff, Jeremy Pinto, Anaum Khan, Louis-François Bouchard, Chenglei Si, et al. Ignore this title and HackAPrompt: Exposing systemic vulnerabilities of LLMs through a global scale prompt hacking competition. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. 137,000+ adversarial interactions, 44 defenses all eventually bypassed.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy (SP)*, pages 3–18, 2017. doi: 10.1109/SP.2017.41.

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 (ICLR)*, 2014. ICLR 2024 Test of Time Award.

Tencent Keen Security Lab. Experimental security research of Tesla autopilot. https://keenlab.tencent.com/en/2019/03/29/

Tencent-Keen-Security-Lab-Experimental-Security-Research-of-Tesla-Autopilot/, 2019. Published March 29, 2019.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction APIs. In *25th USENIX Security Symposium*, pages 601–618, 2016.

Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. DolphinAttack: Inaudible voice commands. In *ACM Conference on Computer and Communications Security (CCS)*, pages 103–117, 2017. doi: 10.1145/3133956.3134052. ACM CCS 2017 Best Paper Award.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. Spotlight paper; arXiv:2307.15043, July 27, 2023.