A Meta-Analysis of Machine Learning Security Research: Attack-Defense Dynamics, Technical Evolution, and Cross-Concept Patterns

Anonymous authors
Paper under double-blind review

Abstract

Machine learning (ML) security research has grown rapidly, yet systematic understanding of its technical evolution, attack-defense dynamics, and cross-concept patterns remains limited. This study presents a systematic meta-analysis of 1,591 security papers spanning six security topics and five ML concepts, released between January 1, 2018, and June 30, 2024. Beyond analyzing research trends, we quantify the attack-defense imbalance across all concept-topic combinations, revealing that defense research significantly lags behind attack research in emerging areas such as LLM jailbreaks (defense ratio = 0.30) and text-to-image membership inference (0.20). Using LLM-assisted annotation of all paper abstracts, we identify 32 distinct technique families and trace their evolution over time, finding that attack techniques such as backdoor injection and adversarial perturbation first appeared in federated learning and graph neural networks before being adopted in LLM and text-to-image model security research. We further identify 17 technique families shared across multiple ML concepts, with six spanning all five concepts studied. Additionally, we examine factors associated with academic influence, finding that ML concepts, security topics, author count, regions, collaboration patterns, and publication status are all statistically significantly associated with citation density. Our findings highlight critical defense gaps, map the technical landscape of ML security, and suggest concrete directions for future research.

1 Introduction

Machine learning (ML) has transformed a wide range of industries, such as healthcare Wiens & Shenoy (2018); Binder et al. (2021); Klauschen et al. (2018), finance Goodell et al. (2021); Dixon et al. (2020), chemistry Stocker et al. (2020); Keith et al. (2021), and bioinformatics Senior et al. (2020); ngoc et al. (2020). While the widespread adoption has been transformative, the increasing integration of ML systems into decision-making processes and critical operations has amplified concerns regarding security, fairness, privacy Xu et al. (2021); Al-Rubaie & Chang (2019), and the consequences of real-world incidents ube. Consequently, understanding and addressing the security of ML systems has become a key focus.

Recent studies have highlighted security threats across various ML paradigms, including federated learning, contrastive learning, large language models (LLMs), text-to-image models, etc. These threats span a broad spectrum of attack vectors, such as model stealing attacks that compromise intellectual property Tramèr et al. (2016); Shen et al. (2022); Kariyappa et al. (2021); Sanyal et al. (2022), data poisoning attacks that corrupt the training process Biggio et al. (2012); Chen et al. (2021); Li et al. (2021); Salem et al. (2022), and jailbreak attacks that bypass the security mechanisms of LLMs Wei et al. (2023); Li et al. (2023); Deng et al. (2023); Shen et al. (2024); Chao et al. (2023). The consequences of these attacks are particularly severe in domains where system integrity is paramount, such as autonomous driving systems Deng et al. (2021) and healthcare applications Newaz et al. (2020), where compromised ML models can lead to catastrophic outcomes and significant financial losses.

Such threats have led to substantial research activity in the ML security field, characterized by an ongoing competition between attackers and defenders. Attackers continuously explore novel attack vectors and

methods to circumvent existing defenses, while defenders work to design more robust architectures and develop advanced countermeasures Wu & Wang (2021); Wang et al. (2020); Huang et al. (2022); Chu et al. (2024); Tramèr et al. (2017). This dynamic interplay has resulted in an exponential increase in academic publications within the ML security field, reflecting the complexity of challenges, the diversity of potential solutions, and the ever-changing nature of the research field.

Despite the surge in research activity, the landscape of ML security research remains largely unknown. Although researchers and practitioners acknowledge the general growth trajectory, and numerous surveys on ML security have been published Kuntla et al. (2021); Guan et al. (2018), there remains a notable gap in systematically and quantitatively understanding publication patterns, emerging research directions, and the evolution of specific security concerns. This gap limits effective resource allocation, research prioritization, and strategic planning, potentially hindering efforts to address ML security challenges effectively.

This study aims to fill this gap by conducting a systematic meta-study of ML security papers. Specifically, we focus on five major ML concepts: federated learning, contrastive learning, large language models, text-to-image models, and graph neural networks; and six key security topics, including adversarial examples, data poisoning attacks, model stealing attacks, membership inference attacks, jailbreak attacks, and prompt injection attacks.

Our study focuses on three research questions (RQs):

RQ1: How has the ML security research landscape evolved over time?

RQ2: What technical patterns emerge across ML security research?

RQ3: What factors are associated with the academic influence of ML security papers?

Our Approach. We collect 1,963 papers from Semantic Scholar Sem and DBLP DBL, covering six security topics across five ML concepts, released between January 1, 2018, and June 30, 2024. After manual review by five domain experts, we retain 1,591 papers. For RQ1, we analyze paper distribution and temporal trends and quantify the attack-defense imbalance across all concept-topic combinations. For RQ2, we use LLM-assisted annotation on all paper abstracts to extract technique families, then trace their evolution over time and identify cross-concept shared technique families. For RQ3, we employ hypothesis testing to evaluate whether six factors (ML concepts, security topics, author count, geographic regions, collaboration patterns, and publication status) are associated with academic influence.[1]

Main Findings. The main findings are summarized below, organized by research question.

- The field's growth is uneven: attack research count consistently outpaces defense, and the gap is widest in the fastest-growing areas. For example, LLM jailbreak research has a defense ratio of only 0.30. This suggests that research effort alone does not ensure security readiness (see Section 3).
- We identify 32 distinct families (30 technique families and two analysis/tooling categories). The field is undergoing a structural shift from white-box, training-time settings toward black-box, inference-time settings, where defenses remain scarce. We further find that six technique families span all five ML concepts, revealing that certain security threats (e.g., data poisoning, membership inference) are rooted in the fundamental properties of machine learning rather than in any specific architecture (see Section 4).
- All six factors we examined (ML concepts, security topics, author count, regions, collaboration patterns, and publication status) are statistically significantly associated with citation density. Broader collaboration and timely dissemination (even as preprints) are associated with greater academic reach (see Section 5).[2]

[1]The claim of artifacts and the ethical considerations are addressed in Appendix D and Appendix 10 in the supplementary material, respectively.

[2]We do not believe citation metrics should drive research priorities. However, understanding the factors correlated with academic influence can help researchers make informed decisions about collaboration and dissemination strategies.

Table 1: Five ML concepts studied in this work.

Concept	Abbr.	Description
Federated Learning	FL	A distributed ML training strategy designed to enhance privacy protection McMahan et al. (2016).
Contrastive Learning	CL	An unsupervised learning method that learns data representations by comparing similarities and differences between samples Chen et al. (2020); Coates et al. (2011).
Large Language Model	LLM	A pre-trained model capable of understanding various types of data and generating human-like responses, including both unimodal and multimodal models OpenAI (2023); Yin et al. (2023).
Text-to-Image Model	T2IM	A generative model that produces images from natural language descriptions.
Graph Neural Network	GNN	A neural network designed to process graph-structured data Scarselli et al. (2009).

Table 2: Six security topics studied in this work. Our study covers both attacks and their corresponding defenses.

Topic	Abbr.	Description
Adversarial Example	AE	Inputs subtly perturbed to cause incorrect model predictions Goodfellow et al. (2014).
Data Poisoning	DP	Manipulating training data to alter model behavior; backdoor attacks Chu et al. (2024) are a representative form Biggio et al. (2012).
Model Stealing	MS	Extracting model information (weights, architecture) through query interactions Tramèr et al. (2016).
Membership Inference	MIA	Inferring whether a data point was used in model training Shokri et al. (2017); Choo et al. (2021).
Jailbreak	JB	Manipulating model input to bypass safeguards and elicit prohibited content Wei et al. (2023); Li et al. (2023).
Prompt Injection	PI	Manipulating input to cause the model to deviate from its intended purpose Greshake et al. (2023); Perez & Ribeiro (2022).

2 Research Data

2.1 Research Scope

Machine learning is a rapidly evolving field, with new concepts, spanning algorithms, frameworks, and models, being introduced continuously, leading to consistent improvements in model performance. However, high-performance models are accompanied by growing security and privacy risks Tramèr et al. (2016); Shen et al. (2024); Zhang et al. (2023). Consequently, researchers have devoted significant efforts to studying both attacks on and defenses for these concepts. While we acknowledge that it is impossible to cover all ML concepts and security topics in such a flourishing research domain with limited human resources, our study focuses on those proposed between January 1, 2018, and June 30, 2024, as they are typically prioritized by the recent research community and are more likely to yield statistically significant conclusions. Specifically, our study focuses on papers about six security topics against five ML concepts.

We focus on five ML concepts that are widely used and increasingly exposed to real-world security threats (Table 1), and six security topics covering the major types of attacks and corresponding defense mechanisms Chowdhury et al. (2024); Rosenberg et al. (2021); Paracha et al. (2024) (Table 2).

Note, in this study, we focus more on how the ML concepts are targeted by the security topics, rather than how security topics are implemented using ML concepts. For instance, regarding a paper that uses Contrastive Learning (CL) to generate adversarial examples (AE) against Text-to-Image Models (T2IMs), the ML concept of the paper is "T2IMs" rather than "CL."

2.2 Data Sources

We rely on two academic search engines as the data sources.

Semantic Scholar Sem. This service has a collection of over 200 million publications across a wide range of scientific fields, offering extensive coverage of papers from various publishers and preprint databases.

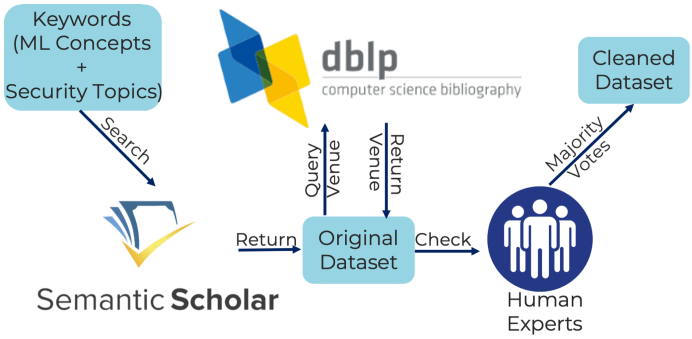


Figure 1: Dataset collection workflow.

Moreover, it supports match query functionality sem, enabling us to obtain related papers by providing corresponding keywords.

DBLP DBL. Certain attributes, such as the publication venue, are occasionally missing from the papers retrieved via the Semantic Scholar API. To address this issue, we integrate the DBLP API to retrieve the missing attributes for these papers. Additional details regarding these two data sources are provided in Appendix B.

2.3 Data Collection

In this section, we first introduce our data collection workflow and then elaborate on each step in the workflow.

Overview. The overall data collection workflow is presented in Figure 1. The process begins with the bulk matching query of paper titles and abstracts, using various keywords through the Semantic Scholar API to identify related publications. We then use the DBLP API to fill in the missing venue attributes of the collected papers. Upon completing these steps, we obtain an original dataset. We then conduct a human review process to filter out papers that do not meet the requirements and augment the data with attributes such as institutes. The paper collection and screening process follows the PRISMA pri guideline, as illustrated in Figure 23 (Appendix G).

Publications Collection via Semantic Scholar API. To systematically collect relevant papers for our study, we first developed keyword sets for the five ML concepts and the six safe topics, respectively. These keyword sets are then used to generate 30 sets of keyword pairs by combining the keyword sets for the five ML concepts with those for the six security topics in a pairwise manner. The details of the keyword set selection process, the entire keyword sets, and the pairing strategy are provided in Appendix A. These keyword pair sets enable us to leverage the bulk-matching query function of the Semantic Scholar API sem to identify relevant papers. The bulk matching query function of the Semantic Scholar API can retrieve relevant papers within a specified time range and support paging to handle large query results. We utilized the Semantic Scholar API to query these 30 sets of keyword pairs sequentially, setting the search period from January 1, 2018, to June 30, 2024, to obtain 30 subdatasets. In each subdataset, we labeled every entry with "MLConcept" and "SecurityTopic" based on the keyword pair used in the query. For instance, in a subdataset obtained using the keyword pair ("LLM," "MS"), entries were labeled with "MLConcept" as "LLM" and "SecurityTopic" as "MS." Together, these 30 subdatasets constituted our original dataset.

Completion of Missing Publication Venues via DBLP API. Among the data obtained from Semantic Scholar, we found 84 papers missing their publication venues. To address this issue, we use DBLP to fill in the missing venues. Specifically, we queried each paper's title to retrieve its publication venue by DBLP Search API DBL and restricted the API to return only the single most relevant result. We then verified the accuracy of the retrieved data by comparing the authors in the DBLP result with the authors listed for the corresponding paper in the original dataset. If the authors matched, we deemed the result accurate and updated the missing venue information accordingly. After processing all 84 papers, we successfully retrieved

Table 3: The distribution of the final cleaned dataset (1,591 unique papers; 1,682 concept×topic entries).

Security Topics \ ML Concepts	Federated Learning	Contrastive Learning	Large Language Model	Text-to-Image Model	Graph Neural Network	**Total**
Adversarial Example (AE)	48	30	55	38	33	204
Data Poisoning Attack (DP)	706	35	82	29	64	**916**
Model Stealing Attack (MS)	76	6	11	3	17	113
Membership Inference Attack (MIA)	111	5	23	20	45	204
Jailbreak Attack (JB)	-	-	200	5	-	205
Prompt Injection Attack (PI)	-	-	40	-	-	40
Total	**941**	76	411	95	159	1,682[1]

[1] Some papers cover multiple topics, thus they are recorded multiple times in the dataset, causing the total number to be higher than 1,591.

publication venues for 13 papers. The remaining 71 papers either lacked venue information in DBLP or were not found in the database. To avoid potential bias introduced by excessive manual intervention, we omit them in the final dataset.

Human Annotation. The original dataset, which comes from Semantic Scholar and is supplemented by DBLP, may contain some papers of low relevance. For instance, a paper on industry security might mention backdoor attacks as a contextual background, even though its primary focus is not on backdoor attacks. Specifically, we engage five human experts, each either pursuing or holding a Ph.D. degree in computer science and having at least one year of professional experience in ML security, to perform human annotation tasks. For each paper, three experts independently review the title and abstract, assigning the labels (i.e., "IfKeep," "MLConcept," "SecurityTopic," "AttackAccess," and "AttackPhase") accordingly. Benchmarks, surveys, and systemization of knowledge papers are excluded from our study. Each paper is annotated three times by different experts, and the final label is determined by majority vote. The annotation demonstrates a substantial inter-agreement among the labelers (Fleiss' Kappa = 0.742) Falotico & Quatto (2015). In addition, human experts manually review each paper to identify affiliated institutes and regions. Our entire human annotation process costs over 300 person-hours.

Data Statistics. Leveraging Semantic Scholar and DBLP, we collected 1,963 security papers released from January 1, 2018, to June 30, 2024, and their related metadata as the original dataset, including the titles, abstracts, venues, available dates, authors, and citation counts, covering six security topics targeting five ML concepts. After human annotation, we finalized the cleaned dataset comprising 1,591 unique papers. The detailed paper distribution of the final cleaned dataset is shown in Table 3. The overview of the attribute of the cleaned dataset is provided in Table 11 in Appendix G.

3 Research Landscape and Attack-Defense Dynamics (RQ1)

In this section, we investigate the distribution of security papers across different concepts and how the focus of security topics targeting different ML concepts has evolved over the past six and a half years. We first analyze the distribution from both static and temporal perspectives, then examine the attack-defense balance across all concept-topic combinations to identify areas where defense research is most urgently needed.

3.1 Paper Distribution and Temporal Trend Across ML Concepts

Overall Results. As shown in Table 3, among all ML concepts we study, federated learning has the largest number of papers, with 941 papers, followed by LLMs with 411 papers. The least studied ML concept is contrastive learning, with only 76 papers, less than 8.3% of federated learning.

Paper Distribution of Five ML Concepts. We find that for different ML concepts, the most popular security topic against them varies, as shown in Table 3. Specifically, data poisoning attacks are the most popular security topic against federated learning, contrastive learning, and graph neural networks, with 706, 35, and 64, respectively. The security papers for data poisoning attacks against the above three ML concepts account for 75.0%, 46.1%, and 40.3% of the total number of security papers against them separately. Regarding LLMs, jailbreak attacks are particularly prominent, with the highest number of security papers

(200), which makes up 48.7% of the total number of security papers against LLMs. For text-to-image models, adversarial examples remain the most studied, with 38 papers representing 40.0% of the total.

Temporal Trends of Five ML Concepts. In Figure 2, we illustrate the temporal trends in the number of new security papers against ML concepts per quarter from January 1, 2018, to June 30, 2024. The number of new security papers per quarter on all five ML concepts has been steadily increasing from January 1, 2018, to June 30, 2024.

Specifically, the number of new security papers related to federated learning per quarter has continuously grown since 2019, with a rapid surge beginning in 2021. From 2018 until the end of 2023, federated learning consistently held the top position in terms of the number of new security papers per quarter. Research on LLMs began gaining significant attention in 2022 and witnessed a sharp surge in the same year. By 2024, the number of new papers in this topic surpassed that of federated learning, making it the most researched ML concept. Meanwhile, the number of new security papers per quarter on text-to-image models, graph neural networks, and contrastive learning has shown steady yet relatively modest growth, with a more gradual upward trend compared to the other two concepts.

Temporal Trends of Six Security Topics against Each ML Concept. Next, we perform a more detailed analysis of how the number of new security papers of six security topics per quarter has evolved over time on each ML concept, with the results shown in Figure 3 and Figure 24 in Appendix G.

Federated Learning. Figure 2b shows the trend in the number of new security papers of federated learning from 2018 to 2024, categorized by four types of attacks: adversarial examples, backdoor attacks, model stealing, and membership inference attacks. Data poisoning attacks dominated the number of new security papers per quarter, starting to rise significantly around 2021 and accelerating sharply after 2022. By 2024, the number of new papers per quarter in this topic has exceeded 83, far exceeding the other topics. The other three security topics, i.e., adversarial examples, model stealing, and MIA, had relatively stable and small numbers of new papers, fluctuating between 0 and 12 papers over the years. While MIA had occasional spikes, it was still close to the level of adversarial examples and model stealing. None of the three methods has seen the same explosion as poisoning and backdoor attacks. These data suggested that backdoor attacks were clearly gaining attention in the context of federated learning, especially in the last few years.

LLMs. Figure 2a illustrates the trends in the number of new security papers focusing on LLMs per quarter from 2018 to the first half of 2024. Jailbreak attacks exhibit the most significant growth, particularly starting in 2023. In the second quarter of 2024, the number of new jailbreak papers against LLMs per quarter reached 79, far exceeding papers on other security topics. Data poisoning attacks also experienced notable growth, becoming the second most discussed topic by 2024, with 28 papers. The other four attack methods, i.e., adversarial examples, model stealing, MIA, and prompt injection, showed an increase in the number of new security papers per quarter starting in 2023, but their growth was relatively moderate compared to the sharp rise seen in jailbreak and backdoor attacks.

GNN, CL, and T2IMs. The trends of GNN, CL, and T2IMs are illustrated in Figure 24 in Appendix G. Overall, we observed that GNN demonstrated a similar trend to that of federated learning. For contrastive learning and text-to-image models, the number of new security papers per quarter on various security topics began to increase in 2020 and 2022, respectively. However, due to the relatively small number of papers, no clear trends can be observed.

In summary, federated learning consistently dominated security research while LLMs surged a lot. This is potentially driven by data privacy regulations like GDPR GDP, which came into effect in 2018. However, driven by ChatGPT's public release and subsequent widespread adoption, LLM security papers have surged in 2023, overtaking federated learning in Q1 2024 (see Figure 2). This surge also indirectly reflects the popularity of LLMs themselves.

3.2 Paper Distribution and Temporal Trend Across Security Topics

Overall Results. As shown in Table 3, among all the security topics, papers related to data poisoning attacks are the most numerous, totaling 916. Adversarial examples, membership inference attacks, and jailbreak attacks each have a similar number of papers, around 200. Model-stealing attack papers are

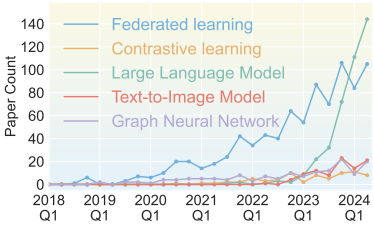
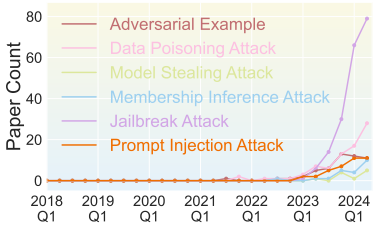
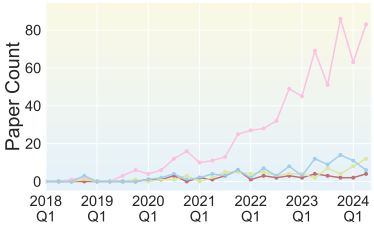


Figure 2: The number of new security papers against ML concepts. Each data point refers to the total number of new papers in a quarter.



(a) LLMs



(b) Federated Learning

Figure 3: The number of new papers on selected ML concepts (LLMs and FL) per quarter. Each data point refers to the total number of new papers in a quarter.

relatively fewer, with only 113, despite spanning all five ML concepts. Additionally, prompt injection attack papers have the smallest amount, with only 40, as this topic is exclusively associated with LLMs.

Paper Distribution of Six Security Topics. From Table 3, it could be easily found that for data poisoning attacks, model stealing, and membership inference attacks, federated learning emerged as the most popular topic. The number of papers for data poisoning attacks, model stealing, and membership inference attacks against federated learning was 706, 76, and 111, accounting for 77.1%, 67.3%, and 54.4% of the total, respectively. In addition, federated learning had a dominant position in these three security topics, and its number of papers far exceeded that of other ML concepts. In contrast, the most popular ML concept for the remaining three attack methods (adversarial examples, jailbreak, and prompt injection) was LLMs, with 55, 200, and 40.

Temporal Trend of Six Security Topics. Figure 4 shows the overall trends in the number of new security papers per quarter related to six security topics. The number of new papers per quarter on data poisoning attacks has experienced the most significant growth over time, with a noticeable increase starting around 2020. By the second quarter of 2024, it reached the highest number, exceeding 132. From the second quarter of 2019 to the second quarter of 2024, it has been the most popular security topic, with a much higher number of papers than other security topics. Research on jailbreak attacks began to surge rapidly in 2023, and by 2024, this security topic had become the second most studied attack method, with the number of new papers per quarter surpassing 82 from the first quarter of 2024. The number of new papers per quarter on adversarial examples, model stealing, and MIA has shown steady but relatively slower growth over the years. As for prompt injection, this type showed a slight increase in the number of new security papers per quarter starting in 2023, but compared to poisoning and jailbreak attacks, it remains one of the less explored topics.

Temporal Trends of Each Security Topic against Five ML Concepts. We also performed a more fine-grained analysis of how the number of new security papers of each ML concept per quarter has evolved over time on each security topic. The related results are shown in Figure 5 and Figure 25. These figures reveal which ML concepts have become increasingly popular under each security topic over time.

AE. First, Figure 4a illustrates the trend of the number of new security papers focusing on adversarial examples against five ML concepts per quarter from 2018 to 2024. The number of new security papers per quarter related to adversarial examples against LLMs began to gain traction in 2022 and show rapid growth thereafter, becoming the most researched topic from the second quarter of 2023. Research on adversarial examples against text-to-image models began to increase significantly around 2023 and has been on a steady upward trend since then. The number of new security papers per quarter on adversarial examples in federated learning started appearing sporadically in 2020 and reached a significant peak of activity around 2021 to 2023. As for the remaining two ML concepts to adversarial examples, both showed limited but sustained research activity with occasional spikes, but both remain relatively less explored compared to LLMs and text-to-image models.

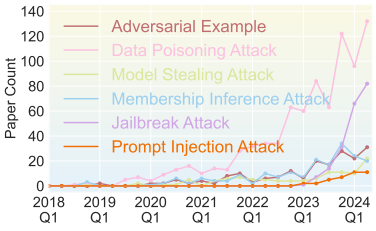
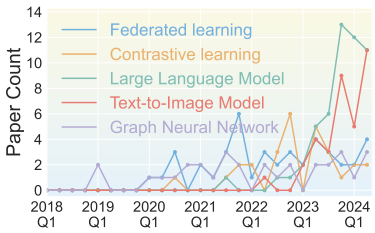
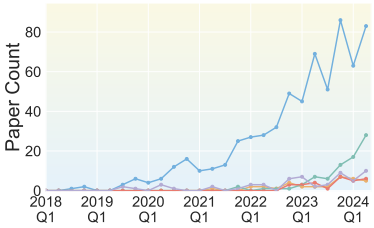


Figure 4: The number of new security papers on security topics. Each data point refers to the total number of new papers in a quarter.



(a) Adversarial Examples



(b) Data Poisoning Attack

Figure 5: The number of new papers on selected security topics (AE and DP) against each ML concept per quarter. Each data point refers to the total number of new papers in a quarter.

DP. In the context of the number of new security papers per quarter for data poisoning attacks from 2018 to 2024, federated learning has experienced rapid growth, particularly since 2021, consistently surpassing other ML concepts and remaining the most prominent topic in the field. LLMs have shown significant growth since 2023, becoming the second most popular topic by the latter half of the year, following federated learning. The other three ML concepts have also seen a gradual increase in attention, but their number of new security papers per quarter remains considerably lower than the top two.

MS and MIA. Similarly, for model stealing and membership inference attacks shown inin Figure 25 in Appendix G, although the number of new papers per quarter was far from that of data poisoning attacks, the overall trend was consistent with data poisoning attacks.

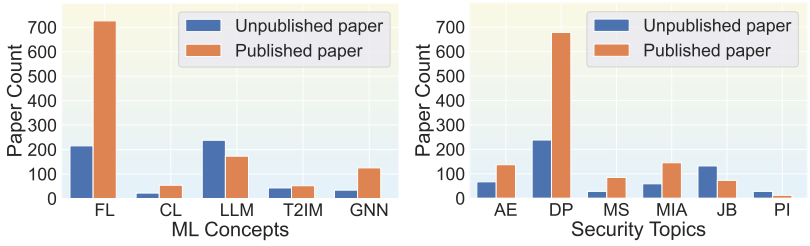
JB and PI. Finally, regarding jailbreak and prompt injection, we found almost only papers against LLMs for both. This also implied the specificity of these two security topics. Despite their differences in number, these two topics exhibited similar trends. Starting in 2023, the number of new security papers per quarter began to rise significantly, reaching a peak in 2024. This indicates that, with the widespread adoption of LLMs, the potential risks posed by jailbreak and prompt injection attacks have increasingly drawn the attention of the academic community.

In summary, data poisoning attacks remain the most researched security topic. Given ML models' reliance on large, uncurated datasets, this highlights the diverse range of attack vectors associated with these threats and their significant real-world implications (e.g., Google incident goo). Security research on federated learning predominantly focuses on data poisoning attacks, whereas studies on LLMs concentrate on jailbreak attacks.

3.3 Paper Distribution of Publication Statuses

Finally, we analyze which ML concepts of security papers are more likely to be published, and which security topics have higher publication rates for security papers. The publication rate in our study is defined as the percentage of published papers on a topic relative to the total number of papers on the same topic between January 1, 2018, and June 30, 2024. And we defined preprints as unpublished papers and papers published in scientific journals or conferences as published papers. First, with respect to which ML concepts attack and defense papers are more likely to be published, Figure 6a shows that contrastive learning has the highest publication rate (0.786) among the five ML concepts, followed by GNNs. Federated learning also has a high publication rate of 0.772, but slightly lower than contrastive learning. Large language models are the only ML concept where the number of unpublished papers exceeds the number of published papers.

Figure 6b illustrates the comparison of the number of unpublished papers to the number of published papers against the six attack methods. Among these, model stealing has the highest rate of published papers (0.752). Following this, data poisoning ranks second, with a slightly lower publication rate of 0.740. In contrast, jailbreak and prompt injection are the only two attack methods where the number of unpublished papers exceeds the number of published papers, with published rates of 0.356 and 0.300, respectively.



(a) ML concepts (b) Security topics

Figure 6: The number of published versus unpublished papers.

LLM security papers, despite their high volume, show the lowest publication rate, particularly in jailbreak and prompt injection attacks. One possible explanation is that the current work is still in its early stages and has not been sufficiently refined, or that the vulnerabilities studied have been quickly fixed, leading to the community's lack of widespread acceptance that it is sufficient for publication. However, this interpretation remains speculative and warrants further investigation.

3.4 Attack-Defense Imbalance

Beyond counting the total number of papers, we analyze the balance between attack and defense research across all concept-topic combinations. For each combination, we compute the *defense ratio*, defined as the number of defense papers divided by the total number of papers in that combination. Papers labeled as "Both" (containing both attack and defense contributions) are counted toward both categories.

Overall Results. Figure 7 presents the defense ratio for all 30 concept-topic combinations. Several findings stand out. First, federated learning exhibits relatively balanced attack-defense research, with defense ratios above 0.50 for most security topics (e.g., FL×DP = 0.74, FL×MS = 0.87). Second, LLM security research is significantly skewed toward attacks: LLM×JB has a defense ratio of only 0.30, LLM×DP = 0.31, and LLM×MIA = 0.26. Third, similar imbalances appear for text-to-image models (T2IM×MIA = 0.20, T2IM×DP = 0.28) and graph neural networks (GNN×DP = 0.30, GNN×AE = 0.36).

In summary, a systematic attack-defense imbalance exists in ML security research. Emerging areas (LLMs, T2IMs) have defense ratios below 0.40, while mature areas (FL) have ratios above 0.55.

Defense Lags. We also quantify the *defense lag*, defined as the number of quarters between the first attack paper and the first defense paper for each combination. Table 18 (Appendix G) summarizes the defense lag for all non-empty concept-topic combinations. Notable positive defense lags, where defense trails attack, include CL×DP (7 quarters), GNN×MS (8 quarters), LLM×DP (5 quarters), and T2IM×JB (4 quarters), indicating that defense research often takes substantial time to respond to newly emerging attack vectors.

Several combinations, however, exhibit *negative* defense lags, meaning that papers categorized as defenses appeared before the corresponding attack papers. Rather than suggesting that defense systematically outpaced attack, this pattern indicates that many protective mechanisms were initially developed under broader or adjacent security objectives and were only later associated with a more clearly defined attack category. For example, in FL×MS (lag = −13), differential privacy and secure aggregation were introduced as general privacy protections for federated learning as early as 2018, well before model stealing attacks against FL were explicitly studied in 2022. A similar pattern appears in LLM×MS (lag = −2), where watermarking methods for copyright protection preceded explicit studies of model extraction attacks. More broadly, this finding suggests that effective defenses often emerge not from reacting to a single attack formulation, but from developing transferable protection mechanisms that remain useful across evolving and partially overlapping threat boundaries. This observation therefore highlights the value of proactive defense research and encourages the community to design more generalizable defenses, rather than focusing exclusively on attack-specific responses.

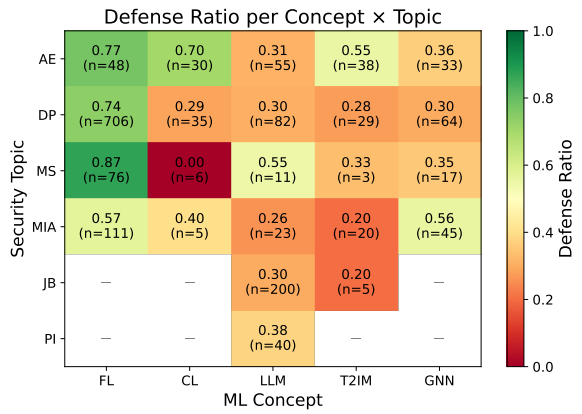


Figure 7: Defense ratio for each concept×topic combination. Green indicates balanced research; red indicates attack-dominated areas.

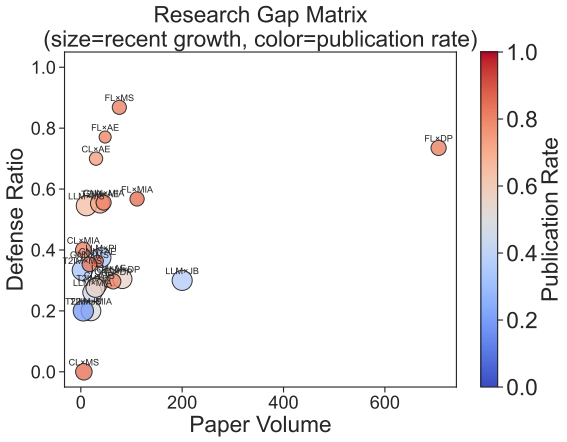


Figure 8: Research opportunity map. Each point represents a concept×topic combination. x-axis: paper volume; y-axis: defense ratio; point size: recent growth (2023–2024); color: publication rate.

Potential Research Opportunities. Figure 8 provides an integrated view by plotting each concept-topic combination along four dimensions: paper volume, defense ratio, recent growth rate, and publication rate. This research opportunity map reveals three distinct regions: (1) mature combinations with high volume and balanced attack-defense research (e.g., FL×DP); (2) rapidly growing but defense-deficient combinations (e.g., LLM×JB); and (3) sparsely studied combinations that represent potential research gaps (e.g., CL×MS with a defense ratio = 0, T2IM×MS with only 3 papers).

These patterns suggest that future work should prioritize not only publication volume, but also the structural imbalance between attack development and defense maturity across combinations. In particular, the second region appears especially promising for high-impact research, because rapid growth coupled with low defense coverage indicates that threat awareness is advancing faster than defensive understanding. For these combinations, important next steps include developing more realistic evaluation protocols and more generalizable defense mechanisms rather than isolated point solutions. By contrast, the first region is less likely to benefit from further incremental gains alone and may instead require consolidation-oriented efforts such as standardized benchmarks, reproducibility studies, and deployment-aware evaluation. Meanwhile, the third region offers opportunities for problem-defining research, where the establishment of clear attack surfaces, realistic application scenarios, and initial baselines may be as valuable as proposing new methods.

3.5 Takeaways

First, defense research should be prioritized in areas where the attack-defense gap is widest. The community's current research effort is unevenly distributed. Some combinations have accumulated hundreds of attack papers with minimal defensive counterparts. Directing new research toward these under-defended areas would yield greater security improvement per paper than further strengthening already-balanced areas. **Second, proactive defense design is more effective than reactive response.** Our defense lag analysis shows that broadly designed protection mechanisms can provide coverage before specific attacks are formally defined. Rather than waiting for each new attack to appear and then developing a tailored defense, the community should invest in generalizable defense frameworks that remain effective across evolving threat boundaries. **Third, the ongoing shift from training-time to inference-time threat models demands new defensive thinking.** The techniques, assumptions, and evaluation norms established in the FL-dominated era are rooted in white-box, training-time settings. As the field moves toward black-box, inference-time scenarios driven by LLM deployment, researchers should not simply adapt existing training-time defenses but develop fundamentally new approaches suited to inference-time constraints.

Table 4: Attack-defense statistics by threat model category. Papers addressing "both" settings are counted in both corresponding quadrants. DefR = defense papers / total papers.

Quadrant	Total	Attack	Defense	DefR
White-box × Training	777	294	473	0.61
Gray-box × Training	117	36	78	0.67
White-box × Inference	136	72	55	0.40
Black-box × Training	74	45	28	0.38
Black-box × Inference	372	279	80	**0.22**

4 Technical Evolution and Cross-Concept Patterns (RQ2)

To move beyond paper-level metadata analysis, we investigate the technical details employed in the 1,591 papers, such as the threat model and the technique families. Since a paper may be relevant to multiple concept×topic combinations, the dataset contains 1,682 entries. Unless otherwise noted, the counts reported in this section refer to unique papers (1,591 total). Within specific concept×topic combinations (e.g., FL×DP), counts refer to entries in that combination.

4.1 Threat Model Landscape and Temporal Shift

Landscape. The threat model assumed by each paper, such as the *access level* (white-box, black-box, or gray-box) and the *attack phase* (training-time or inference-time), reveals fundamental structural patterns in ML security research. Of the 1,591 papers, 893 (56.1%) assume a white-box setting, 439 (27.6%) black-box, and 126 (7.9%) gray-box, with 149 (9.4%) classified as not applicable (e.g., defense-only papers where access level is not defined). Regarding the attack phase, 1,098 (69.0%) target the training phase and 517 (32.5%) target inference time. Papers addressing both settings (16 for access, 31 for phase) are counted in both corresponding categories; 7 papers have no applicable phase. Combining these two dimensions yields a five-cell view (Table 4), revealing two dominant research clusters with very different attack-defense characteristics:

- **Cluster A (training-time, white/gray-box)** encompasses 894 papers (777 white-box + 117 gray-box) with a defense ratio of 0.62. This cluster is dominated by FL security research and features a mature defense ecosystem with five major defense families: anomaly detection, robust aggregation, differential privacy, adversarial training, and federated secure aggregation. Gray-box papers (126 total) are almost exclusively FL-related (94%), reflecting FL's unique threat model where a participating client has white-box access to its local model but limited information about the global model.
- **Cluster B (inference-time, black-box)** contains 372 papers with a defense ratio of only 0.22. This cluster is dominated by LLM security research and features 13 distinct attack families but only two major defense families: safety filters and input preprocessing. The attack-to-defense imbalance in this cluster is the most severe in the dataset.

Temporal Shift. The balance between these two clusters has shifted dramatically over the study period. As shown in Figure 9(a), white-box research declined from 65% of all papers in 2022 to 50% in 2024, while black-box research grew from 10% to 41%. Gray-box research, which is almost exclusively associated with FL (94% of gray-box papers), declined from 18% in 2018–2019 to 5% in 2024 as LLM and T2IM research grew and diluted its share. Similarly, Figure 9(b) shows that training-time research declined from 90% in 2022 to 53% in 2024, while inference-time research grew from 11% to 47%. This shift coincides with the rise of LLM security research: LLMs are predominantly accessed as black-box APIs, and security concerns such as jailbreaks and prompt injection arise at inference time.

The transition has important implications: the defense techniques developed for Cluster A (robust aggregation, anomaly detection of malicious training updates) are largely inapplicable to Cluster B, where defenders can only intervene at the input/output level. As shown in Figure 10(a), the defense ratio in white-box

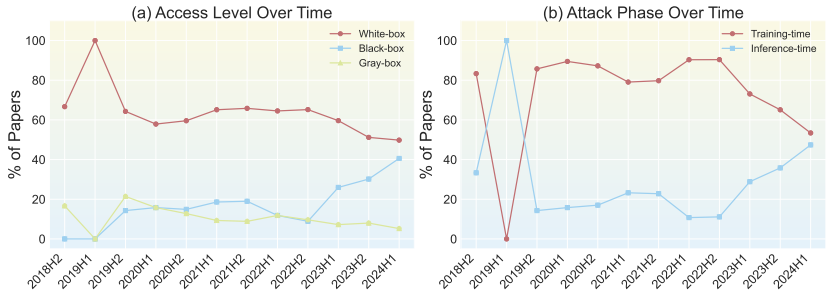


Figure 9: Temporal shift in threat model distribution. (a) Access level: white-box research is declining while black-box research is growing rapidly; gray-box remains a niche FL-specific setting. (b) Attack phase: training-time research is giving way to inference-time research. Papers addressing "both" settings are counted in both categories.

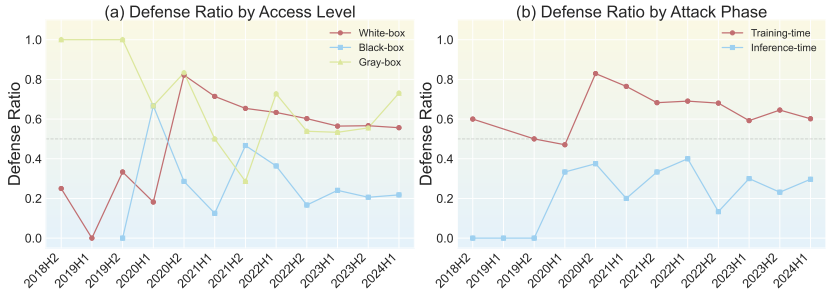


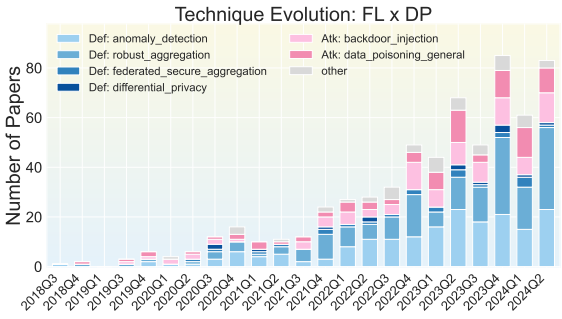
Figure 10: Defense ratio over time by threat model dimension. The dashed gray line marks 0.50 (parity between attack and defense). (a) By access level: white-box and gray-box maintain healthy defense ratios, while black-box remains critically low. (b) By the attack phase, training-time defense is stable, inference-time defense shows slow recovery but remains far from parity.

settings has remained stable at 0.55–0.70, while the defense ratio in black-box settings has stagnated at 0.20–0.25 despite rapid growth in research volume (from 22 papers in 2022 to 195 in 2024). Gray-box settings maintain a relatively high defense ratio (0.60–0.70), reflecting the maturity of FL's defense ecosystem. Figure 10(b) shows a similar pattern by attack phase: training-time defense ratio remains stable around 0.60, while inference-time defense ratio shows a modest recovery from 0.23 in 2022 to 0.29 in 2024, but remains far below parity.

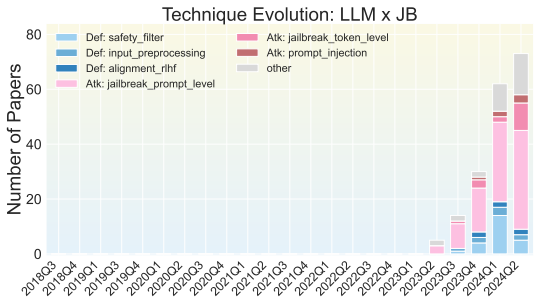
4.2 Technique Family Distribution and Evolution

We introduce the notion of a *technique family*, which groups papers by the specific attack or defense methodology they employ. For example, within the broad category of "data poisoning attacks," we distinguish between *backdoor injection* (planting a hidden trigger during training), *general data poisoning* (corrupting training data without a trigger), and *model inversion* (reconstructing training data from model outputs). Similarly, on the defense side, we distinguish between *robust aggregation* (Byzantine-tolerant aggregation in federated learning), *anomaly detection* (detecting poisoned data or malicious clients), *differential privacy* (adding noise for privacy guarantees), and others. This finer-grained classification enables us to track how specific technical approaches have evolved over time and which approaches are shared across ML concepts.

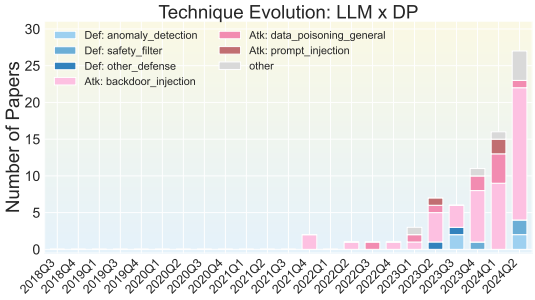
To assign technique families at scale, we use LLM-assisted annotation (Claude Haiku 4.5) on all paper abstracts. For each paper, we extract the primary technique family from a predefined taxonomy of 30 categories, as well as coarse-grained threat model dimensions (white-box vs. black-box, training-time vs. inference-time). The annotation quality was verified through human review, achieving an accuracy above 93%. Details of the annotation taxonomy, prompt, and quality report are provided in Appendix E.



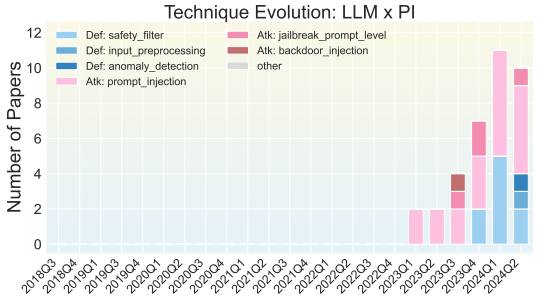
(a) FL × Data Poisoning (697 entries)



(b) LLM × Jailbreak (187 entries)



(c) LLM × Data Poisoning (78 entries)



(d) LLM × Prompt Injection (38 entries)

Figure 11: Technique family evolution over time for four representative concept×topic combinations. Each color represents a technique family; the y-axis shows the number of papers per quarter.

4.2.1 Overall Technique Family Distribution

Across the entire dataset, we identify 32 distinct families (30 technique families and two other families), with the top 10 technique families summarized in Table 5. The distribution reveals that *defense-oriented* technique families collectively outnumber attack-oriented ones. The top three are anomaly detection (234 papers), robust aggregation (204), and backdoor injection (197), reflecting the dominance of data poisoning as a security topic (see Section 3.2) and the maturity of FL's defense ecosystem.

4.2.2 Technique Family Evolution

For each concept-topic combination with at least 20 papers, we trace the temporal evolution of technique families over time, yielding 17 evolution charts in total. To better highlight broader ecosystem-level patterns, we organize the representative cases into two groups: federated learning (FL), which generally exhibits more mature and balanced attack-defense development, and large language models (LLMs), where rapid growth is often accompanied by stronger concentration and weaker defensive diversification. The remaining charts are provided in Appendix G.

FL-Related Combinations. The FL-related combinations generally reflect a more mature stage of development, with broader technical diversity and clearer attack-defense co-evolution. As shown in Figure 11a, FL×DP is one of the most technically mature combinations, comprising 697 entries and 18 technique families. Its defense landscape is dominated by anomaly detection (207 papers) and robust aggregation (200 papers), both of which have grown steadily since 2019, while backdoor injection (102 papers) and general data poisoning (92 papers) constitute the two main attack paradigms. Federated secure aggregation (29 papers), including cryptographic approaches such as secure multi-party computation and homomorphic encryption, further represents a smaller but steadily growing defense direction. A similarly mature pattern appears in FL×MIA, which contains 110 entries and 15 technique families. In this combination, membership inference attacks (36 papers) and differential privacy defenses (35 papers) form a near-symmetric pair, suggesting a

relatively balanced line of research in which attack and defense have evolved in tandem. Property inference (9 papers) broadens the attack landscape as a related but distinct threat variant.

LLM-Related Combinations. In contrast, the LLM-related combinations tend to show rapid expansion but less balanced internal development. As illustrated in Figure 11b, LLM×JB is the largest and most diverse of the LLM settings, with 187 entries and 19 technique families, yet the literature remains heavily concentrated around prompt-level jailbreaks, which alone account for 96 papers and 48% of the total. Defensive work remains comparatively limited, with safety filters (24 papers) being the most common countermeasure, while token-level jailbreaks (16 papers), including GCG-style gradient-based attacks, emerged in late 2023 as a distinct attack paradigm. LLM×DP exhibits a similarly unbalanced structure, though in a different form: backdoor injection dominates the combination with 46 out of 78 entries (59%), whereas general data poisoning (11 entries) and anomaly detection (4 entries) remain much less developed, indicating that defensive work is still in an early stage. Finally, LLM×PI represents a relatively young and narrow area, with only 38 entries and 6 technique families in total. The literature is currently organized around a simple attack-defense structure, with prompt injection attacks (22 papers) opposed mainly by safety filters (9 papers), suggesting that this area is still in an early phase of conceptual and technical expansion.

Cross-Combination Observations. Comparing across combinations, we observe three broad maturity patterns. FL×DP represents the most mature and diversified setting, with 18 technique families, a relatively balanced distribution between attack and defense research, and multiple competing defense paradigms. By contrast, LLM×JB appears to be growing but still concentrated: although it also contains 19 technique families, the literature is dominated by a single attack paradigm, namely prompt-level jailbreaks, while defensive work remains comparatively limited. In comparison, combinations such as LLM×PI and T2IM×AE remain nascent, with fewer than 10 technique families and little diversity in defense strategies, suggesting that these areas are still at an early stage of development.

4.3 Cross-Concept Shared Technique Families

We analyze which technique families are shared across multiple ML concepts. A technique family is considered "shared" if it appears in at least two ML concepts with three or more unique papers in each.

We identify 17 shared technique families (Figure 12 and Table 6). Six technique families span all five ML concepts: anomaly detection, backdoor injection, general data poisoning, membership inference, adversarial perturbation, and input preprocessing. These results demonstrate that certain attack/defense approaches represent fundamental security concerns that transcend specific ML architectures.

Technique Family Diffusion Patterns. We further trace the *temporal adoption order* of shared technique families, defined as the sequence in which a technique family first emerges across different ML concepts. A consistent pattern can be observed: many technique families appear first in federated learning or graph neural networks, and are later adopted in contrastive learning, LLMs, and text-to-image models. This trend is illustrated by several representative cases. For anomaly detection, the earliest appearances follow FL (2018 Q3) → GNN (2019 Q3) → LLM (2023 Q3) → T2IM (2023 Q3) → CL (2023 Q4). Backdoor injection exhibits a similar trajectory, emerging in FL (2019 Q3), then GNN (2020 Q2), CL (2021 Q2), LLM (2021 Q4), and finally T2IM (2022 Q4). Membership inference follows the same overall pattern, first appearing in FL (2018 Q4), then GNN (2020 Q4), CL (2021 Q1), LLM (2022 Q3), and T2IM (2022 Q4).

This FL/GNN-first pattern holds for 7 of the top 10 shared technique families. The exceptions are adversarial perturbation (GNN first), transfer attack (CL first), and watermarking (GNN first). This suggests that FL and GNN, as the earlier and more established ML concepts in security research, serve as early adopters for security technique families that are later explored in newer concepts.

Attack vs. Defense Paradigm Sharing. Among the 17 shared technique families, 9 are primarily attack-oriented and 8 are primarily defense-oriented. However, defense technique families tend to have relatively lower cross-concept coverage: 3 defense families (anomaly detection, input preprocessing, adversarial training) span 4 or more concepts, compared to 5 attack families (backdoor injection, data poisoning, membership inference, adversarial perturbation, transfer attack). This suggests that while attack ideas transfer relatively freely across ML concepts, defense solutions may be more concept-specific.

Table 5: Top 10 technique families by unique paper count. Two non-technique categories (empirical analysis and toolkit) are excluded. The complete taxonomy of 32 families is in Appendix E.

Technique Family	#Papers	Atk/Def
Anomaly detection	234	Defense
Robust aggregation	204	Defense
Backdoor injection	197	Attack
Data poisoning (general)	142	Attack
Jailbreak (prompt-level)	98	Attack
Membership inference	90	Attack
Adversarial training	79	Defense
Differential privacy	77	Defense
Adversarial perturbation	60	Attack
Input preprocessing	59	Defense

Table 6: Top 10 cross-concept shared technique families, ranked by concept coverage. "Papers" counts each unique paper once. "First Appearance" shows the earliest concept-quarter pair.

Technique Family	#Concepts	#Papers	First Appearance
Anomaly detection	5	234	FL (2018Q3)
Backdoor injection	5	197	FL (2019Q3)
Data poisoning (gen.)	5	142	FL (2018Q4)
Membership inference	5	90	FL (2018Q4)
Adversarial training	4	79	GNN (2019Q3)
Differential privacy	3	77	FL (2018Q4)
Adversarial perturbation	5	60	GNN (2019Q1)
Input preprocessing	5	59	FL (2021Q2)
Watermarking	4	18	GNN (2021Q4)
Transfer attack	4	17	CL (2022Q4)

Why Certain Technique Families Are Shared Across Concepts. The cross-concept recurrence of certain technique families is not coincidental; it reflects underlying structural commonalities among ML systems. We identify three levels of shareability based on what the paradigm exploits or protects.

The most broadly shared technique families exploit *inherent properties of ML* that are independent of any specific architecture. Data poisoning (5 concepts, 142 unique papers) exploits the universal dependence of ML models on training data: regardless of whether the target is an FL global model, a GNN classifier, or a fine-tuned LLM, corrupting training data can alter model behavior. Membership inference (5 concepts, 90 unique papers) exploits the tendency of models to memorize training data, a property observed across all ML paradigms. Notably, membership inference is the only shared technique family that operates predominantly in a black-box, inference-time setting across *all* concepts (e.g., 0/18 white-box in LLM, 2/15 in T2IM, 2/17 in GNN), indicating that its attack surface is truly independent of model internals. Backdoor injection (5 concepts, 197 unique papers) exploits the fact that any model trained via gradient optimization can have hidden behaviors implanted during training, with a highly consistent white-box, training-time threat model across concepts.

A second level of sharing arises from *commonalities in learning paradigms*. Adversarial training (4 concepts, 79 unique papers) applies the min-max optimization framework to any differentiable model, with consistent white-box, training-time threat models across FL, CL, LLM, and GNN. Its absence in T2IM may reflect the difficulty of adapting adversarial training to the denoising process of diffusion models. Transfer attacks (4 concepts, 17 unique papers) exploit shared representation spaces in models that use pre-trained features, with a fully consistent black-box, inference-time threat model.

A third level involves *convergent functional needs*: different concepts independently develop similar techniques to address analogous requirements. Watermarking (4 concepts, 18 unique papers) addresses intellectual property protection across FL, LLM, T2IM, and GNN, but the implementations diverge substantially—FL embeds watermarks in model parameters, while LLM and T2IM embed them in outputs—resulting in inconsistent threat models across concepts.

In contrast, technique families that remain concept-specific typically depend on *architecture-specific features*. Robust aggregation (204 papers) and federated secure aggregation (52 papers) are exclusive to FL because they rely on the distributed client-server aggregation architecture that other concepts lack. Similarly, prompt-level jailbreaks (98 papers), token-level jailbreaks (17 papers), prompt injection (29 papers), and safety filters (35 papers) are exclusive to LLM because they depend on natural language interfaces and instruction-following capabilities absent in FL, CL, and GNN. Among the six concept-specific technique families, three belong to FL and three to LLM, reflecting how these two most-researched concepts have each developed unique security ecosystems that the other cannot leverage.

4.4 Implications for Building Secure ML Systems

The analyses in this section yield several practical implications. **First, security risks for new ML concepts can be anticipated before attacks emerge.** Paradigms rooted in inherent ML properties

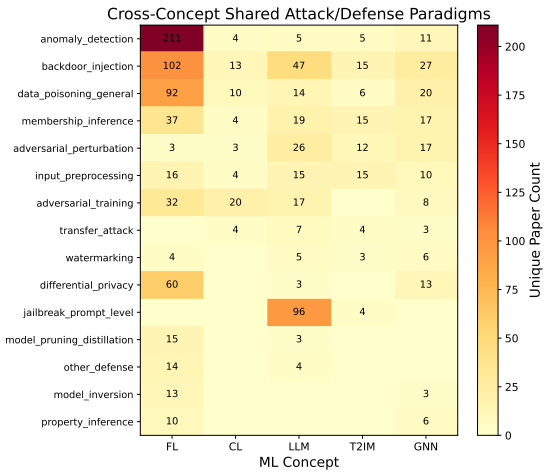


Figure 12: Cross-concept shared technique families. Each cell shows the number of unique papers employing a given technique family within an ML concept. Only families appearing in 2+ concepts with ≥3 papers each are shown.

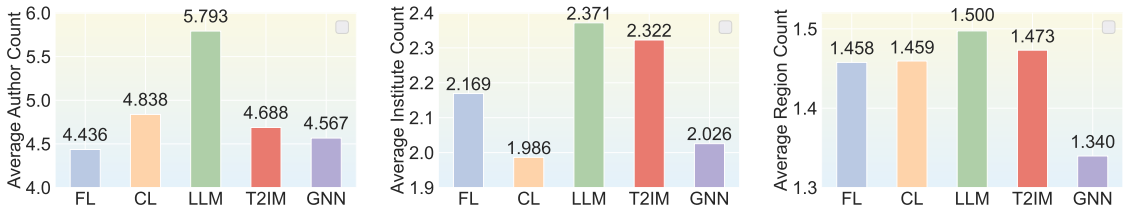
appeared in all five concepts we studied. Any future ML paradigm that relies on training data and gradient optimization will face the same threats. Researchers should begin developing defenses for data poisoning, backdoor injection, and membership inference as soon as a new concept gains traction, rather than waiting for attacks to be demonstrated. **Second, black-box inference-time settings urgently need native defense paradigms.** The current defense ecosystem contains five mature paradigms for white-box training-time settings but only two for black-box inference-time settings (safety filters and input preprocessing). As ML systems move from research prototypes to deployed services, this gap will become increasingly critical. New defense approaches that do not assume access to model parameters or the ability to retrain are worth exploring. **Third, the community should build cross-concept defense infrastructure for universal threats.** Universal attack paradigms are currently met with concept-specific defenses developed in isolation. Shared benchmarks, evaluation protocols, and reusable toolkits would reduce duplicated effort and help defenses mature faster. **Fourth, architecture-specific threats should be identified at design time, not after deployment.** The LLM experience shows that discovering threats only after wide adoption leads to a prolonged defense gap. When designing new architectures, practitioners should proactively analyze what unique attack surfaces the architecture introduces and invest in early defense research.

5 Academic Attributions, Collaboration, and Influence Factors (RQ3)

In this section, we investigate the distribution and temporal trend of academic attributions and the corresponding collaboration patterns. Specifically, we study the distribution and temporal evolution of concentration trends in terms of author count, institutes, and regions, and the proportion of different collaboration patterns. Mean and median are two commonly used statistical measures to describe the central tendency of data. The mean reflects the overall trend of the data, while the median is more suitable for describing the typical value, especially when the data distribution is skewed or contains outliers Khorana et al. (2023). In this section, we primarily use the mean to analyze the overall trend of academic attributions, as the mean considers all data points and intuitively represents the overall level. The results regarding the median will be specifically presented in Section 5.4 to further reveal the typical level of the data and address potential distribution skewness.

5.1 Academic Attributions

Authors. Figure 13a illustrates the average author count per security paper across five ML concepts. We observed that papers on the topic of LLMs have a significantly higher average author count compared to



(a) Average Author Count (b) Average Institute Count (c) Average Region Count

Figure 13: The Average Author Count, Institute Count, and Region Count for Security Papers of five ML Concepts.

Table 7: Top six regions in terms of number of papers.

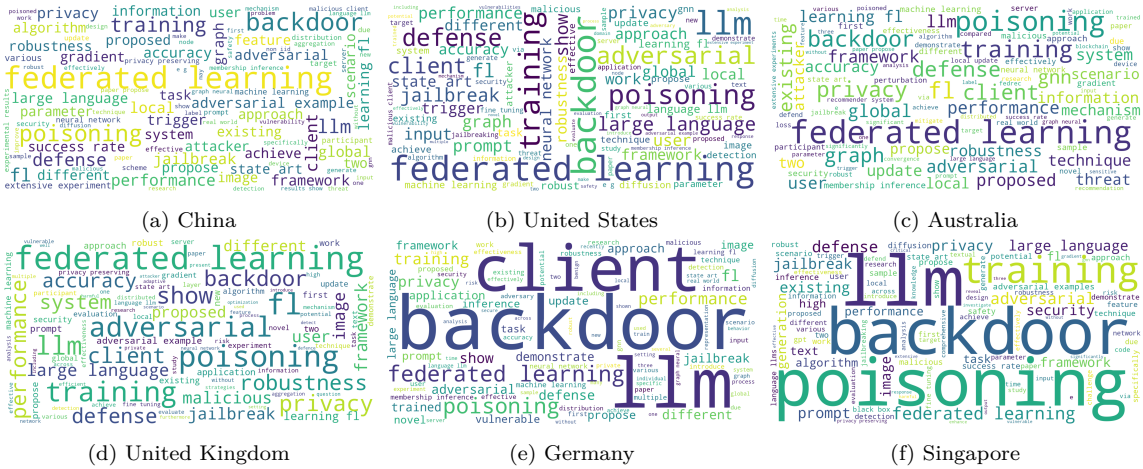
Region	CN	USA	AUS	UK	DE	SGP
# of papers	739	571	120	109	98	82

the other four topics. Specifically, the average author count for LLMs papers exceeds that of the second-ranked contrastive learning by 0.955. Federated learning papers have the fewest average authors; however, the difference was relatively small. Excluding LLMs, federated learning papers have only up to 0.402 fewer authors on average compared to other topics. From a temporal perspective, the average author count per paper has generally increased from 4.000 in 2018 to 5.251 in 2024 (see Figure 26a in Appendix G). However, the average author number of each ML concept does not follow a consistent trend. Notably, since 2021, the average author count for LLMs security papers has always been the highest each year.

Affiliated Institutes. In Figure 13b, we present the average affiliated institute numbers of security papers for each ML concept. We observed that similar to the author count, the number of affiliated institutes for security papers related to LLMs is the highest among the five ML concepts, exceeding the lowest (CL) by 0.381. This may be attributed to the nature of LLM research, which often requires more intensive computation and fosters greater collaboration between academia and industry. Additionally, another computationally intensive topic, T2IM, also exhibits a relatively high number of affiliated institutes, ranking second with a mean of 2.322. On the other hand, regarding the temporal change in affiliated institute number, no consistent trend is observed (see Figure 26b in Appendix G). Overall, the number of affiliated institutes remains relatively stable throughout the period we studied.

Affiliated Regions. The average numbers of affiliated regions are shown in Figure 13c. The overall results are similar to those observed for institutes, with one notable exception: CL. Although CL has the fewest affiliated institutes (ranking fifth), it does not have the fewest affiliated regions, ranking third in this aspect. Also, we do not observe a consistent temporal trend in the number of regions affiliated with the papers (see Figure 26c in Appendix G).

In addition, we compile the top six regions with the most security papers, as shown in Table 7. China and the United States rank first and second, with 739 and 571 papers, respectively. We observe that ML security research is being widely conducted globally, with the top six regions distributed across three different continents. The word cloud of the abstracts and titles of security papers from the six regions is visualized in Figure 14. We find that "federated learning" is the most prominent term in the security papers from China. On the other hand, the hot spots in the United States are more diverse, with "federated learning," "llm" (including "large language"), "poisoning" (including "backdoor"), and "client" all being popular terms. Security papers focused on LLMs exhibit significantly higher average numbers of authors, affiliated institutions, and regions compared to those on the other four ML concepts. This may suggest that security research on LLMs may attract a broader range of researchers from diverse institutions and countries, highlighting the global interest and collaborative nature of this field.



(a) China (b) United States (c) Australia

(d) United Kingdom (e) Germany (f) Singapore

Figure 14: The word clouds for the top six regions with ML security papers.

5.2 Collaboration Patterns

We categorize collaboration patterns based on the number of affiliated institutes and regions associated with each paper into three types: independent work, domestic collaboration, and international collaboration. Independent work refers to papers completed either by a single institute or an individual (institute count = 1, region count = 1). Domestic collaboration involves papers jointly completed by multiple institutes, all located within the same region (institute count > 1, region count = 1). International collaboration refers to papers co-authored by multiple institutes, with at least two of them belonging to different regions (institute count > 1, region count > 1).

The nested pie chart in Figure 27 in Appendix G illustrates the proportion of papers corresponding to the three collaboration patterns. Specifically, 545 papers were completed independently, accounting for 34.4% of the total. The number of papers involving domestic collaboration is 529, representing 33.4% of the total, while those resulting from international collaboration are 510, making up 32.2%. The distribution of papers across the three collaboration patterns is relatively balanced.

The number of papers completed through independent efforts, domestic collaboration, and international collaboration is evenly distributed. This shows that the research collaboration patterns in this field are flexible, and the field demonstrates both openness and a globalized nature.

5.3 Publication Venue Distribution

We classify publication venues into four categories: security venues (e.g., S&P, CCS, USENIX Security, NDSS, TIFS), ML/AI venues (e.g., NeurIPS, ICML, ICLR, CVPR, ACL), other venues, and unpublished/preprints (primarily arXiv).

The venue distribution varies substantially across ML concepts. LLM security papers have the highest unpublished rate (55.4%) and the lowest security venue rate (5.2%), consistent with the preprint-first publication culture in the LLM community. In contrast, FL and GNN security papers have lower unpublished rates (19.3% and 19.7%, respectively) and are more likely to appear in security venues (11.7% and 15.3%). CL security papers have the highest ML/AI venue rate (31.1%), reflecting the closer alignment of contrastive learning with the ML community.

These venue patterns, shown in Figure 15, suggest that ML security research straddles two communities—security and ML—with the balance varying by ML concept. LLM security research is more strongly anchored in the ML community, while FL and GNN security research has deeper roots in the security community. The full venue classification rules are provided in Appendix E.

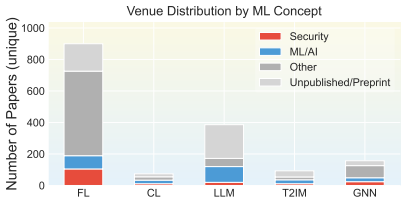


Figure 15: Publication venue distribution by ML concept. Venues are classified into security conferences/journals, ML/AI conferences/journals, other venues, and unpublished preprints.

5.4 Potential Factors Affecting Academic Influence

The academic influence of a paper refers to the extent to which the paper has had an impact on the academic community, which can be affected by a combination of factors. In this section, we investigate whether the academic influence is associated with six different factors, i.e., ML concepts, security topics, author count, regions, collaboration patterns, and publication status, through hypothesis testing.

5.4.1 Experimental Settings

Academic Influence. The academic influence of a paper is defined as the degree to which it has impacted the academic community. This influence can be quantified using two metrics: citation count and citation density (i.e., the average number of citations per day since the paper's publication). Previous studies have demonstrated that both metrics serve as effective indicators of a publication's academic influence Jones et al. (2017); Kadic et al. (2020); Sandison (1975). Given that citation density removes the effects of time, we employ this metric to assess the academic influence of a paper.

Key Factors. Through hypothesis testing, we explore the relationship between academic influence (i.e., citation density) and six key factors: ML concepts, security topics, author count, geographic regions, collaboration patterns, and publication status. Furthermore, we investigate the specific conditions within each factor that correlate with stronger academic influence.

Sample Size. For the hypothesis testing in this section, we analyze at the unique paper level rather than the entry level used in Sections 3–4. Since some papers appear in multiple concept-topic combinations, we deduplicate by paper ID. Of the 1,591 unique papers, 1,569 have sufficient metadata (publication date and citation count) to compute citation density; the remaining 22 papers are excluded from the hypothesis testing due to missing dates. As shown in the supplementary tables in Appendix G, the sample sizes for the ML concept (x_1) and security topic (x_2) analyses are 1,569 (sum of all group counts). For other variables (x_3–x_6), sample sizes range from 1,475 to 1,569 due to additional missing metadata in some fields.

5.4.2 Statistical Methodology

Overview. In this section, we perform comprehensive statistical tests to explore the relationships between academic influence, measured as citation density (dependent variable), and six independent variables. Details of the statistical tests conducted in this study are presented in Appendix C. For each test, we begin by examining the citation density distributions within each group of the independent variables. Based on the characteristics of these distributions, we select appropriate hypothesis tests to assess whether statistically significant differences exist among the groups. Subsequently, we conduct post-hoc analyses to derive fine-grained insights. We aim to assess if the specific group pairs exhibit statistically significant differences and understand the extent of these differences in citation density distributions. In the following, we provide a detailed explanation of each step in our assessment.

Variables of Statistical Tests. We use the following variable settings for our statistical tests:

- **Dependent Variables:** y = citation density.

- **Independent Variables:** x_1 = ML concepts, x_2 = security topics, x_3 = author count, x_4 = regions, x_5 = collaboration patterns, and x_6 = publication status.

Distributional Tests. In this study, for each independent variable, we perform a one-sample Kolmogorov-Smirnov test on each group to verify whether the data follows a normal distribution. The test compares the empirical cumulative distribution function (ECDF) of the sample with the cumulative distribution function (CDF) of a theoretical normal distribution, and determines whether the data follows a normal distribution. The null hypothesis assumes that the data follows a normal distribution. We use the p-value to determine whether the data within each group of the independent variable follows a normal distribution. If the p-value is less than the significance level ($\alpha = 0.05$), the null hypothesis is rejected, indicating that the data in that group does not follow a normal distribution. The results of this test guide the selection of the appropriate subsequent statistical analysis.

Overall Difference Test. We use hypothesis testing to examine whether there are statistically significant differences in citation density between the different groups of an independent variable, thereby assessing whether there is a statistical relationship between the independent and dependent variables. Based on the results of the normality test, we determine whether to use parametric or non-parametric tests for hypothesis testing. Since the dependent variable is numerical, if the normality test indicates that each group follows a normal distribution, we will use parametric tests, such as the t-test Kim (2015) for comparing two groups or ANOVA Scheffé (1999) for comparing more than two groups. If the normality assumption is violated, we will use non-parametric tests, such as the Mann-Whitney U test McKnight & Najab (2010); Nachar et al. (2008) for comparing two independent groups or the Kruskal-Wallis H test McKight & Najab (2010); Ostertagová et al. (2014) for comparing three or more groups. The hypotheses are formulated as follows:

$$H_{0i} : \text{The distribution of } y \text{ is identical across groups of } x_i.$$
$$H_{1i} : \text{The distribution of } y \text{ varies by different groups of } x_i.$$

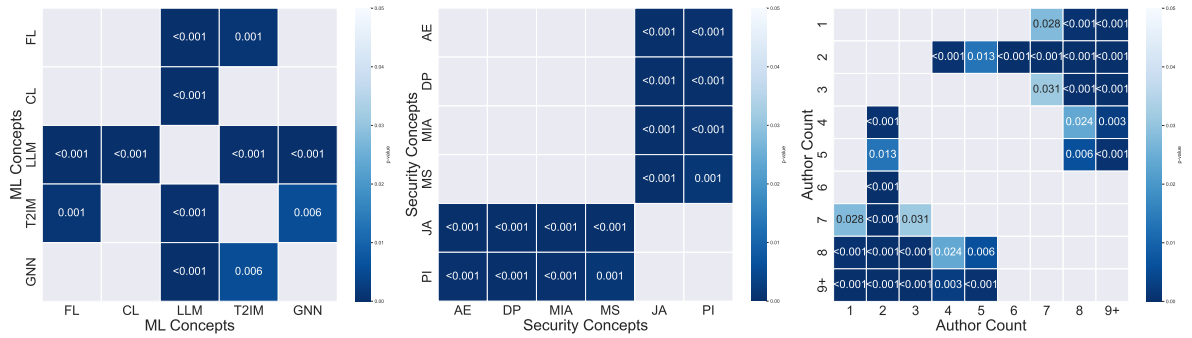
where $i \in \{1, \ldots, 6\}$. We use the p-value to determine whether the data within each group of the independent variable follows a normal distribution. If the p-value is less than the significance level ($\alpha = 0.05$), the null hypothesis is rejected, indicating that there are statistically significant differences in the distribution y between the different groups of x_i.

Post-Hoc Tests. After conducting an overall difference test, such as the Kruskal-Wallis H test, we use Dunn's test Dinno (2015); Pohlert (2014) for post-hoc analysis to further investigate which specific groups exhibit statistically significant differences. The null hypothesis of Dunn's test assumes that the two groups being compared have the same distribution. Dunn's test provides a p-value for each pairwise comparison. To reduce the Type I error rate Banerjee et al. (2009), we apply Bonferroni correction Armstrong (2014) to the p-values to determine whether there are statistically significant differences between the groups. While Bonferroni correction primarily addresses the risk of false positives, it can also provide a more reliable framework for interpreting results when sample sizes are unbalanced. If the p-value is less than the significance level ($\alpha = 0.05$), the null hypothesis is rejected, indicating that there are statistically significant differences in y between these two groups of x_i. Afterwards, by combining the results of Dunn's test and analyzing the median for each group presented in the boxplot of x_i, we can determine under what specific conditions certain categories tend to exhibit higher citation density, complementing the statistical findings from Dunn's test.

Why Focus on Medians. The median is a robust measure of central tendency that is less influenced by outliers and variability. This makes it a more reliable measure when data is not symmetrically distributed (e.g., skewed publication distributions) or contains extreme values (e.g., highly cited publications) Khorana et al. (2023). In such cases, the median provides a more accurate representation of the typical value within a dataset.

5.4.3 Results

The distribution test results show that the p-values for all groups across all independent variables are less than the significance level. The highest p-value, 0.049, is observed for the group with the author count



(a) ML Concept (b) Security Topic (c) Author Count

(d) Affiliated Region (e) Collaboration Pattern (f) Publication Status

Figure 16: p-value metrics of Dunn's test for citation density by different independent variables. Those with p-values larger than 0.05 are omitted.

of 1 under the independent variable x_3, which is still below 0.05. This indicates that the citation density distributions for all groups of all independent variables do not follow a normal distribution. Consequently, we employ non-parametric tests for the next step, the overall difference test. Depending on the number of groups within each independent variable, we use either the Mann-Whitney U test or the Kruskal-Wallis H test.

ML Concept. For the variable x_1, which includes more than two groups, the Kruskal-Wallis H test is performed to assess overall differences. The calculated p-value is less than 0.001, which is significantly lower than the threshold for statistical significance ($\alpha = 0.05$), leading to the rejection of the null hypothesis. The distribution of citation density differs statistically significantly across different ML concepts. This shows that the citation density of a paper is related to its ML concepts.

Further analysis using Dunn's test, combined with the statistics presented in the boxplot, allows for a more detailed examination of which specific groups within x_1 exhibit higher citation density. Each cell of the p-value matrix in Figure 16 shows the p-value obtained from Dunn's test between two groups, and for ease of reading, we only show cells with a p-value less than 0.05. The results from Figure 16a demonstrate that the citation density of papers grouped under LLM differs statistically significantly from all other groups. Considering the median values shown in Figure 17a (with detailed statistics provided in Table 12 in Appendix G), it can be concluded that papers associated with LLM tend to exhibit higher citation densities compared to other ML concepts.

Security papers addressing LLMs are likely to have a greater academic impact than those focused on other machine learning concepts. This indicates that, in the security domain, researchers show greater interest in LLMs than in the other four categories of ML concepts. Similarly, the analysis of security topics reveals that researchers pay more attention to data poisoning and prompt injection than to the other four categories. Notably, prompt injection is an attack specifically targeting LLMs. This further confirms the statistically significant attention that LLMs have received in the security domain.

Security Topic. Regarding the overall difference test, the p-value of the Kruskal-Wallis H test on x_2 is below 0.001, which is well below the statistical significance level ($\alpha = 0.05$). Thus, the null hypothesis is rejected. This indicates that the security topics of a paper may affect its citation density.

The results of Dunn's test indicate statistically significant differences in citation density distributions for papers on jailbreak when compared to those on adversarial examples, jailbreak attacks, membership inference attacks, and model stealing, as shown in Figure 16b. Similarly, papers on prompt injection also exhibit statistically significant differences in citation density compared to the same four security topics. Moreover, the median values presented in Figure 17b (with detailed statistics provided in Table 13 in Appendix G) further support this finding. Consequently, we conclude that papers addressing security topics like jailbreak and prompt injection tend to have higher citation densities than those on other topics.

Author Count. When categorizing the paper data based on the author count, we find that papers with more than 9 authors are rare. For example, only six papers have 11 authors, and just one paper has 12 authors. To avoid sparse data and to ensure more balanced sample sizes across the groups, we combine papers with 9 or more authors into the same group, labeling them as "9+ authors."

The Kruskal-Wallis H test for overall differences on x_3 yields a p-value below 0.001, well below the statistical significance level ($\alpha = 0.05$). The null hypothesis is rejected. This suggests that the citation density on a paper is associated with its author count.

Regarding the post-hoc test on the author count, the p-value metrics in Figure 16c, combined with the median information in Figure 17c (detailed statistics can be found in Table 14 in Appendix G), indicate that papers with more than 8 authors tend to have higher citation densities compared to papers with fewer than 5 authors.

Papers with more than eight authors tend to exhibit stronger academic influence compared to those with fewer than five authors. A higher number of authors increases the likelihood of the paper being a product of cross-institutional or interdisciplinary collaboration, potentially broadening its promotional reach.

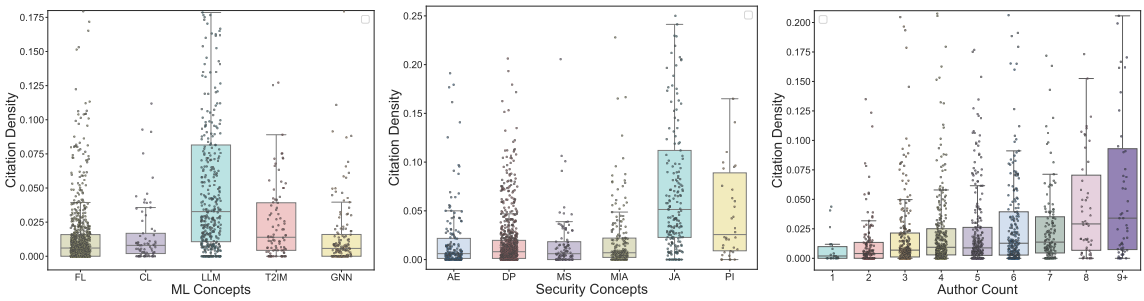
Affiliated Region. In order to prevent the data from being too sparse as well as to ensure the credibility of the statistical results, here we only study regions with more than 30 papers. There are eleven regions with more than 30 papers, in descending order of the number of papers, namely China, the United States, Australia, the United Kingdom, Germany, Singapore, Canada, Japan, South Korea, France, and India.

The Kruskal-Wallis H test for overall differences on x_4 produces a p-value below 0.001, which is significantly lower than the statistical significance ($\alpha = 0.05$). Therefore, the null hypothesis is rejected, indicating that the affiliated region of a paper is associated with its citation density.

The post-hoc test on the affiliated regions, as shown in the p-value metrics in Figure 16d and the median values in Figure 17d (detailed statistics is presented in Table 15 in Appendix G), reveals that the papers affiliated with Singapore tend to have higher citation densities compared to those affiliated with China, Canada, South Korea, France, and India. However, no evidence suggests that papers from any specific region exhibit a statistically significantly higher citation density compared to the citation density of all other regions.

Collaboration Pattern. The Kruskal-Wallis H test for general differences in x_5 produces a p-value below 0.001, significantly lower than the statistical significance level ($\alpha = 0.05$), leading to the rejection of the null hypothesis. The result shows that the collaboration pattern probably influences its citation density. Further analysis of the p-value metrics for collaboration patterns from Dunn's test is shown in Figure 16e, along with the median information in Figure 17e (comprehensive statistical data is provided in Table 16 in Appendix G. This reveals that papers co-authored tend to have higher citation densities compared to those completed independently.

Collaborative papers generally have greater academic influence than those completed independently. Cross-institutional collaborations benefit from broader dissemination channels through the networks of participating institutes, thereby enhancing the academic influence of the paper. To improve the academic influence of their work, authors could consider engaging in cross-institutional collaborations.



(a) ML Concept (b) Security Topic (c) Author Count

(d) Affiliated Region (e) Collaboration Pattern (f) Publication Status

Figure 17: Boxplot for citation density by different independent variables.

Publication Status. In x_6, the data is divided into two groups: published and unpublished. We used the Mann-Whitney U test to conduct an overall difference test to assess whether there is a statistically significant difference between the distributions of the two independent samples. The p-value is 0.036, which is below the statistical significance level ($\alpha = 0.05$), suggesting that the publication status of papers likely influences their citation density.

By observing the differences in medians of different groups in Figure 17f (with detailed statistics shown in Table 17 in Appendix G), we can finally conclude that the citation density of unpublished papers may often be slightly higher than that of published papers.

Unpublished papers tend to have slightly higher citation density than published ones. A potential reason for this could be that some authors of highly influential papers, particularly those from industry, may prioritize starting new projects over investing time in the lengthy submission and peer-review process.

5.5 Summary

All six factors we examined are statistically significantly associated with citation density. Three observations stand out for their practical relevance. **First, topic selection matters more than other factors.** LLM security papers have citation densities several times higher than those on other concepts. This gap is unlikely to reflect quality differences alone. Researchers seeking impact should consider aligning their work with high-attention areas, while also recognizing that under-explored areas offer less competition and potentially higher marginal contribution. **Second, collaboration correlates with influence.** Papers with more authors and cross-institutional or international collaboration tend to receive more citations. While causality cannot be established from our data, this pattern is consistent with the view that diverse teams produce work with broader reach and visibility. **Third, publication venue is not the sole driver of influence.** Unpublished papers have slightly higher citation density than published ones, particularly in fast-moving areas like LLM security. This suggests that timely dissemination through preprints can be as effective as formal publication for accumulating citations, though the two serve different purposes in establishing scientific credibility.

6 Discussion

Defense Gaps as Research Opportunities. Our attack-defense imbalance analysis reveals that defense research is most urgently needed in LLM security, where jailbreak defense ratio is only 0.30 and membership inference defense ratio is 0.26. Similarly, text-to-image model security (T2IM×MIA = 0.20) and GNN security (GNN×DP = 0.30) present significant defense gaps. These imbalances are not merely statistical observations—they represent concrete research opportunities. For instance, while 145 papers have proposed jailbreak attacks against LLMs, only 60 have proposed corresponding defenses, suggesting that the community may benefit from shifting focus toward defense development in these areas.

Cross-Concept Transfer of Security Techniques. Our temporal adoption analysis reveals a consistent pattern: security techniques first appear in federated learning and graph neural networks, and are subsequently explored in LLMs and text-to-image models. This pattern has practical implications. For researchers working on newer ML concepts, techniques developed for earlier paradigms may serve as a starting point. For example, anomaly detection methods for FL could inspire novel defenses against LLM data poisoning. However, the effectiveness of such cross-concept transfer is not guaranteed—the threat models, data modalities, and system architectures differ substantially across concepts. Validating the transferability of defense techniques across ML paradigms is an important open problem.

The LLM Security Paradox. LLM security research exhibits a striking combination of characteristics: the highest paper volume growth, the lowest publication rate (with 55% of papers remaining unpublished), the most severe attack-defense imbalance, and the highest citation density. This suggests a field that is simultaneously the most active, the most impact-generating, and the least mature. The preprint-dominated publication pattern is consistent with the hypothesis that findings in this area may become outdated quickly, as vulnerabilities are frequently patched before papers complete the peer-review cycle, though other explanations (e.g., the novelty of the field, different community norms) are also plausible.

Influence Enhancement. Our findings suggest that papers involving collaborations across multiple regions, institutions, and authors tend to exhibit higher academic impact. This may be attributed to the diversity of perspectives and feedback introduced through collaboration, which can enhance both the quality and influence of the work.

7 Limitations & Future Work

In our study, we made every effort to minimize bias and identify limitations. However, we acknowledge that certain biases and limitations remain unavoidable. First, we collect the original paper dataset by performing relevant searches on the titles and abstracts of the papers with different keywords through the Semantic Scholar API. While we strive to expand the set of keywords for each topic as thoroughly as possible, a small number of less relevant yet related terms may be inadvertently overlooked. This could result in the omission of a few security papers from the original dataset. Second, to ensure the high quality of the final cleaned dataset, we manually annotate the original dataset to identify security papers genuinely relevant to our research topics. However, we also recognize the inherent subjectivity of manual judgment (e.g., different experts may have conflicting opinions on whether a paper should be included). To minimize bias caused by subjectivity, we make each piece of data judged by at least 3 or more experts, and the majority opinion prevails. Third, our findings are limited to the security papers released between January 1, 2018 and June 30, 2024. As new security papers continue to emerge, they may not align with the patterns identified in our study. We leave the analysis of these papers for future research.

8 Related Work

ML Security Surveys. Numerous surveys have been published on specific ML security topics. In adversarial ML, surveys by Biggio and Roli Guan et al. (2018) and Chakraborty et al. Chowdhury et al. (2024) provide comprehensive overviews of adversarial attacks and defenses. For federated learning security, Kairouz et al. Kuntla et al. (2021) survey privacy and security challenges in distributed settings. LLM safety has been surveyed by recent works focusing on jailbreaks, alignment, and prompt injection Rosenberg et al. (2021);

Paracha et al. (2024). However, these surveys are typically scoped to a single ML concept or security topic and rely on narrative synthesis rather than systematic quantitative analysis. Our work differs in two key aspects: (1) we cover *all* combinations of five ML concepts and six security topics simultaneously, enabling cross-concept comparisons that single-topic surveys cannot provide; and (2) we employ quantitative methods (LLM-assisted annotation, statistical hypothesis testing) rather than narrative review, enabling reproducible and scalable analysis.

Meta-Studies of Academic Research. In the broader academic community, meta-studies have been used to understand publication patterns and research dynamics. Fortunato et al. Fortunato et al. (2018) emphasize the importance of meta-studies for improving science policy. Fire et al. Fire & Guestrin (2018) focus on the over-optimization of academic publishing metrics. Smart et al. Smart & Bayer (1986) and Wang et al. Wang et al. (2024) find that international collaboration correlates with higher citations. In computer science specifically, meta-studies have examined artifact availability of Sciences Engineering et al. (2019); Vandewalle et al. (2009); Raghupathi et al. (2022); Collberg & Proebsting (2016), code quality Trisovic et al. (2021), and reproducibility in ML security Olszewski et al. (2023); Hamm et al. (2019). Our work extends this tradition by combining bibliometric analysis with technique-level annotation, bridging the gap between high-level publication statistics and the technical content of the research.

9 Conclusion

In this study, we conducted a data-driven survey of 1,591 papers in the ML security domain, going beyond traditional bibliometric analysis to examine attack-defense dynamics, technical evolution, and cross-concept patterns. Our findings reveal three key contributions. First, we quantify a systematic attack-defense imbalance: defense research significantly lags behind attack research in emerging areas such as LLM jailbreaks and text-to-image model security, while mature areas like federated learning have achieved greater balance. Second, using LLM-assisted annotation of all paper abstracts, we identify 32 technique families and trace their evolution over time, revealing that LLM security remains dominated by a single attack paradigm (prompt-level jailbreaks) with limited defense diversity. Third, we identify 17 technique families shared across multiple ML concepts, with a consistent temporal adoption pattern from federated learning to newer concepts such as LLMs and text-to-image models. Additionally, we confirm statistically significant associations between academic influence and six factors, including ML concepts, security topics, author count, regions, collaboration patterns, and publication status. We hope that our findings, particularly the defense gap analysis and cross-concept paradigm mapping, will help researchers identify high-impact research directions in ML security.

10 Ethical Considerations

We adhere to ethical guidelines and strict data privacy standards, analyzing only publicly available information, such as citation data, publication venues, available dates, and affiliated institutes. Proprietary or sensitive data, as well as harmful data and personal information about authors or developers, are strictly excluded from our analysis.

References

https://dblp.org/search/venue/api.

https://gdpr-info.eu/.

https://www.semanticscholar.org/product/api/.

https://www.un.org/about-us/member-states.

https://support.google.com/translate/thread/45624167/what-if-someone-intentionally-gives-wrong-translation-aka-poisoning-the-model?hl=en.

https://www.prisma-statement.org/.

https://api.semanticscholar.org/api-docs/graph#tag/Paper-Data/operation/get_graph_paper_bulk_search.

https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe.

Mohammad Al-Rubaie and J. Morris Chang. Privacy-Preserving Machine Learning: Threats and Solutions. *IEEE Security & Privacy*, 2019.

Richard A Armstrong. When to use the Bonferroni correction. *Ophthalmic and Physiological Optics*, 2014.

Amitav Banerjee, Udaykumar Bhaskar Chitnis, Sl Jadhav, Js Bhawalkar, and Suprakash Chaudhury. Hypothesis testing, type I and type II errors. *Industrial psychiatry journal*, 2009.

Vance W. Berger and YanYan Zhou. *Kolmogorov–Smirnov Test: Overview*. Wiley Online Library, 2014.

Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning Attacks against Support Vector Machines. In *International Conference on Machine Learning (ICML)*. icml.cc / Omnipress, 2012.

Alexander Binder, Michael Bockmayr, Miriam Hägele, Stephan Wienert, Daniel Heim, Katharina Hellweg, Masaru Ishii, Albrecht Stenzinger, Andreas C. Hocke, Carsten Denkert, Klaus-Robert Müller, and Frederick Klauschen. Morphological and molecular breast cancer profiling through explainable machine learning. *Nature Machine Intelligence*, 2021.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking Black Box Large Language Models in Twenty Queries. *CoRR abs/2310.08419*, 2023.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *International Conference on Machine Learning (ICML)*, pp. 1597–1607. PMLR, 2020.

Xiaoyi Chen, Ahmed Salem, Michael Backes, Shiqing Ma, Qingni Shen, Zhonghai Wu, and Yang Zhang. BadNL: Backdoor Attacks Against NLP Models with Semantic-preserving Improvements. In *Annual Computer Security Applications Conference (ACSAC)*, pp. 554–569. ACSAC, 2021.

Christopher A. Choquette Choo, Florian Tramèr, Nicholas Carlini, and Nicolas Papernot. Label-Only Membership Inference Attacks. In *International Conference on Machine Learning (ICML)*, pp. 1964–1974. PMLR, 2021.

Arijit Ghosh Chowdhury, Md Mofijul Islam, Vaibhav Kumar, Faysal Hossain Shezan, Vaibhav Kumar, Vinija Jain, and Aman Chadha. Breaking down the defenses: A comparative survey of attacks on large language models. *CoRR abs/2403.04786*, 2024.

Junjie Chu, Yugeng Liu, Ziqing Yang, Xinyue Shen, Michael Backes, and Yang Zhang. Comprehensive Assessment of Jailbreak Attacks Against LLMs. *CoRR abs/2402.05668*, 2024.

Adam Coates, Andrew Y. Ng, and Honglak Lee. An Analysis of Single-Layer Networks in Unsupervised Feature Learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 215–223. JMLR, 2011.

Christian Collberg and Todd A. Proebsting. Repeatability in computer systems research. *Communications of the ACM*, 2016.

Gelei Deng, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, and Yang Liu. Jailbreaker: Automated Jailbreak Across Multiple Large Language Model Chatbots. *CoRR abs/2307.08715*, 2023.

Yao Deng, Tiehua Zhang, Guannan Lou, Xi Zheng, Jiong Jin, and Qing-Long Han. Deep Learning-Based Autonomous Driving Systems: A Survey of Attacks and Defenses. In *IEEE Transactions on Industrial Informatics (ITII)*, pp. 7897–7912. IEEE, 2021.

Alexis Dinno. Nonparametric Pairwise Multiple Comparisons in Independent Groups using Dunn's Test. *The Stata Journal*, 2015.

Matthew F Dixon, Igor Halperin, and Paul Bilokon. *Machine learning in finance*. Springer, 2020.

Rosa Falotico and Piero Quatto. Fleiss' kappa statistic without paradoxes. *Quality & Quantity*, 2015.

Michael Fire and Carlos Guestrin. Over-optimization of academic publishing metrics: observing Goodhart's Law in action. *GigaScience*, 2018.

Santo Fortunato, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Stasa Milojevic, Alexander Michael Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, Alessandro Vespignani, Ludo Waltman, Dashun Wang, and Albert Ĺaszló Barabási. Science of Science. *Nature*, 2018.

John W. Goodell, Satish Kumar, Weng Marc Lim, and Debidutta Pattnaik. Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. *Journal of Behavioral and Experimental Finance*, 2021.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. *CoRR abs/1412.6572*, 2014.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. More than you've asked for: A Comprehensive Analysis of Novel Prompt Injection Threats to Application-Integrated Large Language Models. *CoRR abs/2302.12173*, 2023.

Zhenyu Guan, Liangxu Bian, Tao Shang, and Jianwei Liu. When Machine Learning meets Security Issues: A survey. In *IEEE International Conference on Intelligence and Safety for Robotics (ISR)*. IEEE, 2018.

Peter Hamm, David Harborth, and Sebastian Pape. A Systematic Analysis of User Evaluations in Security Research. In *Proceedings of the 14th International Conference on Availability, Reliability and Security*. Association for Computing Machinery, 2019.

Abdelhakim Hannousse. Searching relevant papers for software engineering secondary studies: Semantic Scholar coverage and identification role. *IET Software*, 2021.

Kunzhe Huang, Yiming Li, Baoyuan Wu, Zhan Qin, and Kui Ren. Backdoor Defense via Decoupling the Training Process. *CoRR abs/2202.03423*, 2022.

Richard Jones, Travis Hughes, Kevin Lawson, and Gregory DeSilva. Citation analysis of the 100 most common articles regarding distal radius fractures. *Journal of Clinical Orthopaedics and Trauma*, 2017.

Antonia Jelicic Kadic, Tanja Kovacevic, Edita Runjic, Ana Simicic Majce, Josko Markic, Branka Polic, Julije Mestrovic, and Livia Puljak. Research methodology used in the 50 most cited articles in the field of pediatrics: types of studies that become citation classics. *BMC Medical Research Methodology*, 2020.

Sanjay Kariyappa, Atul Prakash, and Moinuddin K. Qureshi. MAZE: Data-Free Model Stealing Attack Using Zeroth-Order Gradient Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13814–13823. IEEE, 2021.

John A Keith, Valentin Vassilev-Galindo, Bingqing Cheng, Stefan Chmiela, Michael Gastegger, Klaus-Robert Muller, and Alexandre Tkatchenko. Combining machine learning and computational chemistry for predictive insights into chemical systems. *Chemical Reviews*, 2021.

Arjun Khorana, Ayoosh Pareek, Matthieu Ollivier, Sophia J. Madjarova, Kyle N. Kunze, Benedict U. Nwachukwu, Jon Karlsson, Erick M. Marigi, and Riley J. Williams. Choosing the appropriate measure of central tendency: mean, median, or mode? *Knee Surgery, Sports Traumatology, Arthroscopy*, 2023.

Tae Kyun Kim. T test as a parametric statistic. *Korean Journal of Anesthesiology*, 2015.

Frederick Klauschen, Klaus-Robert Müller, Alexander Binder, Michael Bockmayr, Miriam Hägele, Philipp Seegerer, S Wienert, Giancarlo Pruneri, Silvia Teresa De Maria, Sunil S. Badve, Stefan Michiels, Torsten O. Nielsen, Sylvia Adams, Peter Savas, Fraser W. Symmans, Scooter Willis, Tina Gruosso, M. H. Park, Benjamin Haibe-Kains, Brandon D. Gallas, Alastair M. Thompson, Ian A. Cree, Christos Sotiriou, Cinzia Solinas, Matthias Preusser, Stephen M. Hewitt, David. L. Rimm, Giuseppe Viale, Sherene Loi, Sybille Loibl, Roberto Salgado, and Carsten Denkert. Scoring of tumor-infiltrating lymphocytes: From visual estimation to machine learning. *Seminars in Cancer Biology*, 2018.

Gayatri Sravanthi Kuntla, Xin Tian, and Zhigang Li. Security and privacy in machine learning: A survey. *Issues in Information Systems*, 2021.

Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, Jie Huang, and Yangqiu Song. Multi-step Jailbreaking Privacy Attacks on ChatGPT. *CoRR abs/2304.05197*, 2023.

Yiming Li, Tongqing Zhai, Yong Jiang, Zhifeng Li, and Shutao Xia. Backdoor Attack in the Physical World. *CoRR abs/2104.02361*, 2021.

Hubert W. Lilliefors. On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. *Journal of the American Statistical Association*, 1967.

Patrick E. McKight and Julius Najab. Kruskal-wallis test. *The corsini encyclopedia of psychology*, 2010.

Patrick E. McKnight and Julius Najab. Mann–Whitney U Test. *The SAGE Encyclopedia of Research Design*, 2010.

H. B. McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1273–1282, 2016.

Nadim Nachar et al. The mann-whitney u: A test for assessing whether two independent samples come from the same distribution. *Tutorials in quantitative Methods for Psychology*, 2008.

Akm Iqtidar Newaz, Nur Imtiazul Haque, Amit Kumar Sikder, Mohammad Ashiqur Rahman, and Arif Selcuk Uluagac. Adversarial Attacks to Machine Learning-Based Smart Healthcare Systems. In *IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6. IEEE, 2020.

Long Vo ngoc, Cassidy Yunjing Huang, California Jack Cassidy, Claudia Medrano, and James T. Kadonaga. Identification of the Human DPR Promoter Element by using Machine Learning. *Nature*, 2020.

National Academies of Sciences Engineering, Medicine, and Others. *Reproducibility and Replicability in Science*. National Academies Press (US), 2019.

Daniel Olszewski, Allison Lu, Carson Stillman, Kevin Warren, Cole Kitroser, Alejandro Pascual, Divyajyoti Ukirde, Kevin Butler, and Patrick Traynor. "Get in Researchers; We're Measuring Reproducibility": A Reproducibility Study of Machine Learning Papers in Tier 1 Security Conferences. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2023.

OpenAI. GPT-4 Technical Report. *CoRR abs/2303.08774*, 2023.

Eva Ostertagová, Oskar Ostertag, and Jozef Kováč. Methodology and Application of the Kruskal-Wallis Test. *Applied Mechanics and Materials*, 2014.

Anum Paracha, Junaid Arshad, Mohamed Ben Farah, and Khalid Ismail. Machine learning security and privacy: a review of threats and countermeasures. *EURASIP Journal on Information Security*, 2024.

Fábio Perez and Ian Ribeiro. Ignore Previous Prompt: Attack Techniques For Language Models. *CoRR abs/2211.09527*, 2022.

Thorsten Pohlert. The pairwise multiple comparison of mean ranks package. *R package*, 2014.

Wullianallur Raghupathi, Viju Raghupathi, and J. Ren. Reproducibility in computing research: An empirical study. *IEEE Access*, 10:29207–29223, 2022.

Ishai Rosenberg, Asaf Shabtai, Yuval Elovici, and Lior Rokach. Adversarial machine learning attacks and defense methods in the cyber security domain. *ACM Computing Surveys*, 2021.

Graeme Douglas Ruxton and Guy Beauchamp. Time for some a priori thinking about post hoc testing. *Behavioral Ecology*, 2008.

Ahmed Salem, Michael Backes, and Yang Zhang. Get a Model! Model Hijacking Attack Against Machine Learning Models. In *Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2022.

Alexander Sandison. Patterns of citation densities by date of publication in physical review. *Journal of the American Society for Information Science and Technology*, 1975.

Sunandini Sanyal, Sravanti Addepalli, and R. Venkatesh Babu. Towards Data-Free Model Stealing in a Hard Label Setting. *CoRR abs/2204.11022*, 2022.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The Graph Neural Network Model. *IEEE Transactions on Neural Networks*, 2009.

Henry Scheffé. *The Analysis of Variance*. John Wiley & Sons, 1999.

Andrew W. Senior, Richard Evans, John M. Jumper, James Kirkpatrick, L. Sifre, Tim Green, Chongli Qin, Augustin ŽÍdek, Alexander W. R. Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones, David Silver, Koray Kavukcuoglu, and Demis Hassabis. Improved protein structure prediction using potentials from deep learning. *Nature*, 2020.

Xinyue Shen, Yixin Wu, Michael Backes, and Yang Zhang. Voice Jailbreak Attacks Against GPT-4o. *CoRR abs/2405.19103*, 2024.

Yun Shen, Xinlei He, Yufei Han, and Yang Zhang. Model Stealing Attacks Against Inductive Graph Neural Networks. In *IEEE Symposium on Security and Privacy (S&P)*, pp. 1175–1192. IEEE, 2022.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership Inference Attacks Against Machine Learning Models. In *IEEE Symposium on Security and Privacy (S&P)*, pp. 3–18. IEEE, 2017.

John C Smart and Alan E Bayer. Author collaboration and impact: A note on citation rates of single and multiple authored articles. *Scientometrics*, 10:297–305, 1986.

Sina Stocker, Gábor Csányi, Karsten Reuter, and Johannes T. Margraf. Machine learning in chemical reaction space. *Neural Computation*, 2020.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. Stealing Machine Learning Models via Prediction APIs. In *USENIX Security Symposium (USENIX Security)*, pp. 601–618. USENIX, 2016.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble Adversarial Training: Attacks and Defenses. In *International Conference on Learning Representations (ICLR)*, 2017.

Ana Trisovic, Matthew K. Lau, Thomas Pasquier, and Mercè Crosas. A large-scale study on research code quality and execution. *Scientific Data*, 9, 2021.

Patrick Vandewalle, Jelena Kovacevic, and Martin Vetterli. Reproducible research in signal processing. *IEEE Signal Processing Magazine*, 26, 2009.

Jue Wang, Rainer Frietsch, Peter Neuhäusler, and Rosalie Hooi. International collaboration leading to high citations: Global impact or home country effect? *Journal of Informetrics*, 18(4):101565, 2024.

Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations (ICLR)*, 2020.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How Does LLM Safety Training Fail? *CoRR abs/2307.02483*, 2023.

Jenna Wiens and Erica S. Shenoy. Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology. *Clinical Infectious Diseases*, 2018.

Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 16913–16925, 2021.

Runhua Xu, Nathalie Baracaldo, and James B. D. Joshi. Privacy-Preserving Machine Learning: Methods, Challenges and Directions. *CoRR abs/2108.04417*, 2021.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A Survey on Multimodal Large Language Models. *CoRR abs/2306.13549*, 2023.

Boyang Zhang, Xinlei He, Yun Shen, Tianhao Wang, and Yang Zhang. A Plot is Worth a Thousand Words: Model Information Stealing Attacks via Scientific Plots. In *USENIX Security Symposium (USENIX Security)*. USENIX, 2023.

A Keyword Set Selection Process

The keywords for each keyword set were determined through a rigorous and systematic process. Specifically, five experts in the ML security field conducted independent evaluations and engaged in extensive discussions, ultimately reaching a consensus. The final keyword sets of the ML concepts and the security topics are shown in Table 8 and Table 9, respectively. The sets of keyword pairs to collect security papers of a security topic against an ML concept are then constructed as follows:

```
[A keyword set of an ML concept] + [A keyword set of a security topic]
```

Table 8: Keyword sets of ML concepts.

ML Concept	**Keyword Set**
FL	[federated learning, federated reinforcement learning, federated transfer learning]
CL	[contrastive learning]
LLM	[large language models, multimodal models, in-context learning, prompt-based learning, vision language models, multimodal large language models]
T2IM	[text-to-image models, diffusion models, stable diffusion, diffusion-based models, latent diffusion]
GNN	[graph neural networks, graph convolutional networks, graph attention networks, variational graph autoencoders, subgraph neural networks]

B Data Sources

Previous studies have indicated that Semantic Scholar and Google Scholar have similar indexing coverage in the computer science field Hannousse (2021). However, due to the lack of a reliable official API for Google Scholar, we ultimately chose to use Semantic Scholar to obtain the original dataset. Semantic Scholar Sem is a free and powerful academic search engine containing over 200 million papers from all scientific fields and covering papers from different publishers and preprint databases. In this study, we used the Semantic Scholar API to collect metadata for various papers and their corresponding information, including attributes such as citation count, publication venues, authors, and so on.

Table 9: Keyword sets of security topics.

Security Topic	Keyword Set
AE	[adversarial examples, adversarial machine learning, adversarial learning]
DP	[data poisoning, poisoning attacks, trojan attacks, backdoor attack]
MS	[model stealing, model extraction, model piracy, knowledge extraction, model inversion, model-stealing, hyper-parameter stealing, hyperparameter stealing, parameter stealing]
MIA	[membership inference attacks, property inference attacks, attribute inference attacks, inference attacks]
JB	[jailbreak]
PI	[prompt injection]

However, while obtaining data from Semantic Scholar, we found that some data were missing certain attributes, such as publication venue. To address this, we used DBLP to retrieve the missing attributes. DBLP DBL is a well-known online database that provides bibliographic information for major computer science publications. It includes a large number of research papers, conference proceedings, and journal papers from computer science and related fields. DBLP also indexes works from many publishers and preprint databases and provides detailed information about authors, titles, publication venues, and other relevant metadata.

C Details of the statistical tests in our studies

The Kolmogorov-Smirnov (KS) test. is a widely employed non-parametric statistical test used to evaluate whether a sample is drawn from a specific distribution, with its most common application being the assessment of normality Berger & Zhou (2014); Lilliefors (1967).

The Mann-Whitney U test. is a non-parametric test used to assess whether there are statistically significant differences between the distributions of two independent samples, particularly when the data does not follow a normal distribution.

The Kruskal-Wallis H test. is a non-parametric test used to assess whether there are statistically significant differences in the distributions of three or more independent samples. It is an extension of the Mann-Whitney U test, designed to handle differences between multiple groups, and does not require the assumption of normality in the data.

Dunn's test. is a non-parametric post-hoc test used for pairwise comparisons between multiple groups after a statistically significant result from the Kruskal-Wallis H test Ruxton & Beauchamp (2008). It helps identify which specific groups differ from each other when there is evidence that at least one group's distribution is different by comparing the mean ranks of the groups.

D Claim of Artifacts

This study exclusively utilizes open-access resources, including publicly available datasets, interfaces (APIs), and established software libraries. Throughout the research process, we did not create or train any new machine learning models. Consequently, there are no proprietary or unpublished artifacts generated in this work that require submission. However, to facilitate future research and ensure reproducibility, our curated and cleaned datasets will be made available to interested parties upon reasonable request following the acceptance of this paper.

Table 10: Complete technique family taxonomy with definitions and paper counts. Families are grouped by type (attack, defense, analysis) and sorted by count within each group.

Technique Family	Definition	Count
Attack Techniques		
backdoor_injection	Plants a hidden trigger in the model during training, causing targeted misclassification when the trigger is present at inference time.	216
data_poisoning_general	Corrupts training data to degrade model performance or alter predictions, without using a specific trigger pattern (non-backdoor).	150
jailbreak_prompt_level	Uses carefully crafted natural language prompts to bypass LLM safety filters and elicit prohibited content.	107
membership_inference	Infers whether a specific data sample was used in the model's training set, posing a privacy threat.	93
adversarial_perturbation	Crafts small, often imperceptible perturbations to inputs that cause the model to produce incorrect predictions at inference time.	63
prompt_injection	Manipulates LLM input prompts to override system instructions and hijack model behavior for unintended purposes.	33
model_inversion	Reconstructs or approximates training data (e.g., images, text) from model outputs or parameters.	20
property_inference	Infers global properties of the training data distribution (e.g., class ratios, sensitive attributes) from model behavior.	19
transfer_attack	Generates adversarial examples on a surrogate model and transfers them to attack a different target model.	18
jailbreak_token_level	Uses token-level optimization (e.g., GCG, AutoDAN) to generate adversarial suffixes that bypass LLM safety mechanisms.	18
model_extraction	Steals model parameters, architecture, or functionality through query interactions with the target model.	17
other_attack	Attack techniques not covered by the above categories.	8
jailbreak_multi_turn	Uses multi-turn conversation strategies to gradually steer LLMs into producing prohibited content.	5
gradient_based_attack	Uses gradient information (e.g., from a white-box model) as the primary mechanism for constructing the attack.	3
Defense Techniques		
anomaly_detection	Detects poisoned data samples, malicious clients (in FL), or other anomalous inputs to filter out threats.	243
robust_aggregation	Byzantine-robust aggregation methods for federated learning that tolerate malicious client updates.	208
differential_privacy	Applies differential privacy mechanisms (e.g., noise addition, gradient clipping) to protect data privacy.	87
adversarial_training	Augments training with adversarial examples to improve model robustness against adversarial perturbations.	81
input_preprocessing	Detects, transforms, or sanitizes potentially adversarial inputs before they reach the model at inference time.	64
federated_secure_aggregation	Uses secure computation techniques (MPC, homomorphic encryption, secret sharing) for privacy-preserving aggregation in FL.	56
safety_filter	Implements input/output filters for LLM safety, including toxicity detection, content moderation, and guardrails.	37
model_pruning_distillation	Uses model pruning or knowledge distillation as a defense mechanism to remove backdoors or reduce attack surfaces.	23
unlearning	Removes the influence of specific data points or concepts from a trained model without full retraining.	19
other_defense	Defense techniques not covered by the above categories.	19
watermarking	Embeds watermarks into model parameters or outputs for ownership verification and intellectual property protection.	18
certified_defense	Provides provable, mathematically guaranteed robustness bounds against specific classes of perturbations.	14
alignment_rlhf	Uses reinforcement learning from human feedback (RLHF) or similar alignment techniques to improve model safety.	6
Analysis & Tooling		
empirical_analysis	Analyzes, compares, or evaluates existing attack/defense methods through experiments, without proposing a substantially new method.	24
toolkit	Provides a toolkit, framework, or platform for conducting attacks, defenses, or red-teaming evaluations.	10

E Annotation and Classification Details

E.1 LLM-Assisted Annotation

Predefined Taxonomy. Prior to LLM-assisted annotation, we established a predefined taxonomy through a structured human review process. Two annotators with expertise in ML security manually reviewed 300 papers in total. To balance coverage and efficiency, we adopted a tiered sampling strategy: for high-volume concept-topic combinations (those with more than 50 papers, including FL×DP, LLM×JB, LLM×DP, and FL×MIA), we randomly sampled 15% of papers for manual verification; for low-volume combinations (those with 20 or fewer papers), we reviewed all papers. Through iterative discussion and consensus building, the annotators refined the category boundaries and definitions, resulting in a taxonomy of 30 technique families.

Annotation Methodology. We use Claude Haiku 4.5 (via Google Vertex AI) to annotate all 1,591 paper abstracts. Each abstract is processed along with its ML concept and security topic labels. The LLM extracts two types of information: (1) the primary *technique family* from a predefined taxonomy of 30 categories, and (2) *threat model dimensions*, including access level (white-box / black-box / gray-box), attack phase (training-time / inference-time), data modality, and model access type (for LLM papers: open model / closed API). We additionally instruct the LLM to fill in "Others" if they believe it does not belong to any of our predefined technique families.

Taxonomy Design. After the LLM-assisted annotation, 34 entities are labeled as "Others." We then conduct a second-round human review, and the 34 entities were classified into two (empirical analysis and toolkit), yielding the final taxonomy of 32 families (Table 10). Attack families include: backdoor injection, general data poisoning, adversarial perturbation, model extraction, membership inference, prompt injection, jailbreak (prompt-level / token-level / multi-turn), transfer attack, model inversion, property inference, and gradient-based attack. Defense families include: robust aggregation, anomaly detection, differential privacy, adversarial training, certified defense, input preprocessing, model pruning/distillation, unlearning, alignment/RLHF, safety filter, watermarking, and federated secure aggregation.

Quality Assurance. On the 300 human-annotated papers, the LLM-assigned technique family matched the human judgment in over 93% of cases. The most common disagreement involved closely related families (e.g., `backdoor_injection` vs. `data_poisoning_general`), where the LLM sometimes chose the broader category. Note that the taxonomy table (Table 10) reports entry-level counts rather than unique paper counts; since some papers appear in multiple concept-topic combinations, entry counts may be slightly higher.

E.2 Venue Classification Rules

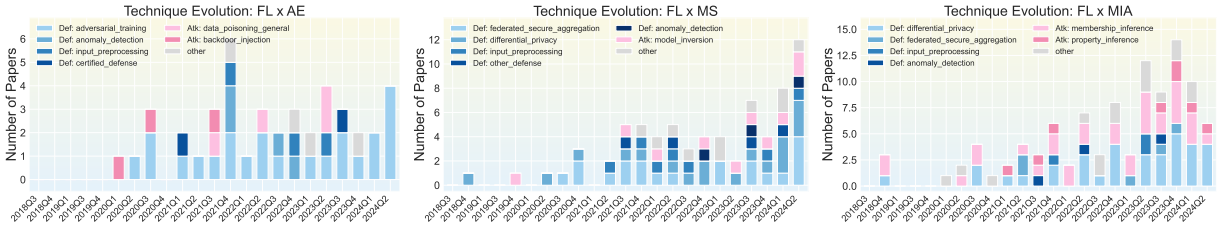
For the venue distribution analysis in Section 5.3, we classify publication venues into four categories using substring matching on venue names (case-insensitive):

- **Security venues**: Venues containing keywords such as "security and privacy," "CCS," "USENIX Security," "NDSS," "ACSAC," "ESORICS," "information forensics and security" (TIFS), "dependable and secure computing" (TDSC), "SaTML," "TrustCom," "PETS," or "RAID."
- **ML/AI venues**: Venues containing keywords such as "NeurIPS," "ICML," "ICLR," "AAAI," "IJCAI," "CVPR," "ICCV," "ECCV," "ACL," "EMNLP," "NAACL," "KDD," "WWW," "JMLR," "TMLR," or "ICASSP."
- **Unpublished/Preprint**: Venues containing "arXiv," "bioRxiv," "OpenReview," or "SSRN."
- **Other**: All remaining venues, including domain-specific journals and workshops not covered above.

Each paper is classified based on its primary venue field from Semantic Scholar. Papers without a venue field are classified as unpublished.

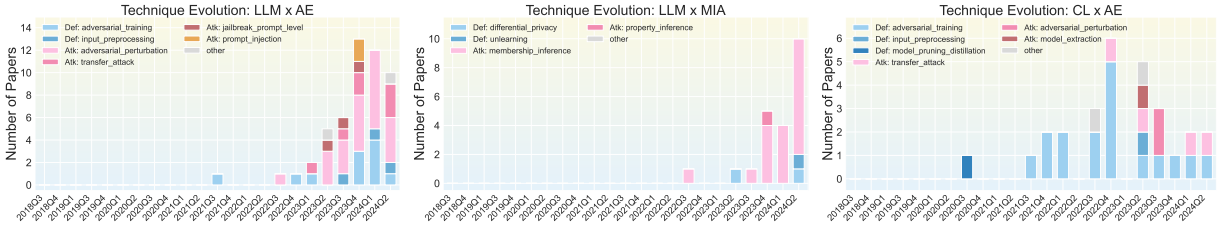
F Supplementary Technique Evolution Figures

This section presents the technique family evolution charts for all 17 concept×topic combinations with at least 20 papers. The four combinations discussed in the main text (FL×DP, LLM×JB, LLM×DP, LLM×PI) are shown in Figure 11. The remaining 13 combinations are shown below.



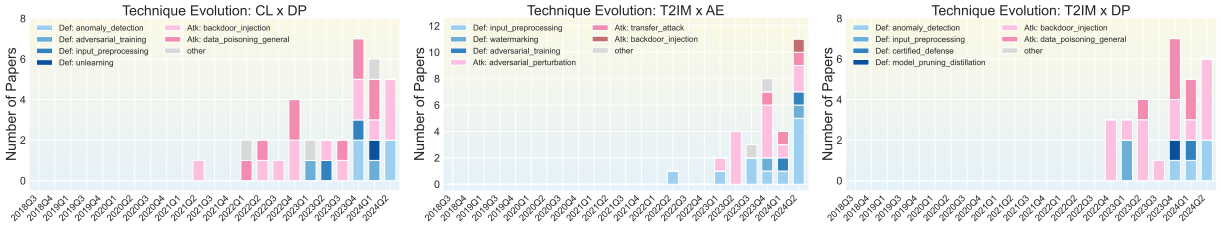
(a) FL × AE (b) FL × MS (c) FL × MIA

Figure 18: Technique evolution for FL (remaining combinations).



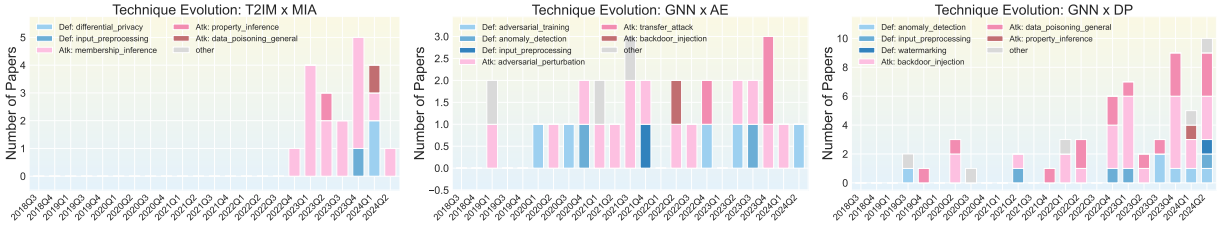
(a) LLM × AE (b) LLM × MIA (c) CL × AE

Figure 19: Technique evolution for LLM (remaining) and CL.



(a) CL × DP (b) T2IM × AE (c) T2IM × DP

Figure 20: Technique evolution for CL (remaining) and T2IM.



(a) T2IM × MIA (b) GNN × AE (c) GNN × DP

Figure 21: Technique evolution for T2IM (remaining) and GNN.

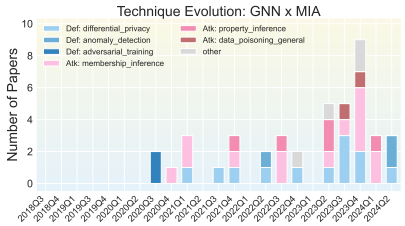


Figure 22: Technique evolution for GNN × MIA.

G Supplementary Tables and Figures

In this section, we present other supplementary figures and tables to provide additional support and clarity for our results.

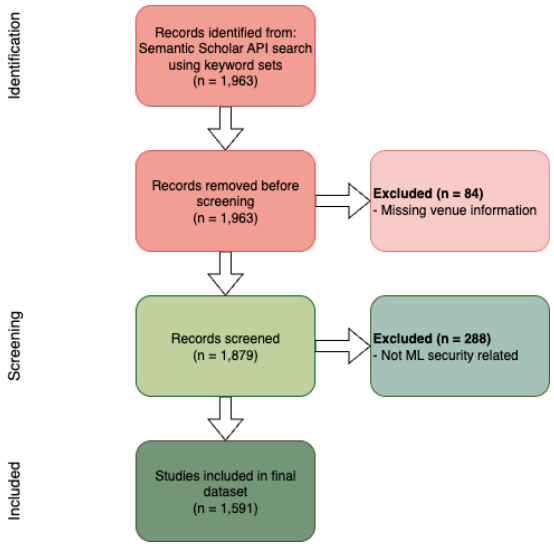


Figure 23: PRISMA flow diagram of the paper collection and screening process.

Table 11: An overview of the cleaned paper dataset's attributes.

Attribute	**Description**	**Data Source**
Title	The title of a paper	Semantic Scholar
Abstract	The abstract of a paper	Semantic Scholar
Venue	The name of the paper's publication venue.	Semantic Scholar & DBLP
Available Date	Date of first public accessibility of a paper	Semantic Scholar
Author	The authors of a paper	Semantic Scholar
Citation Count	Count of times a paper is cited by other papers	Semantic Scholar
Institute	The institute or research unit to which the authors of the paper are affiliated. We focus only on specific universities or research institutes and do not record information about specific faculties or departments.	Human annotation
Region	The region where the institute is located.	Derived from "Institute" with UN.
ML Concept	The ML concept targeted by a paper. Its value is in [*FL*, *CL*, *LLM*, *T2IM*, *GNN*].	Human annotation
Security Topic	The security topic addressed in a paper. Its value is in [*AE*, *DB*, *MS*, *MIA*, *JB*, *PI*].	Human annotation
Type	The main type of a paper. Its value is in [*attack*, *defense*, *both*].	Human annotation

Table 12: Summary statistics of citation density of security papers against five ML concepts.

ML concept	Count	Mean	STD	Min	25%	50%	75%	Max
FL	858	0.019	0.049	0.000	0.000	0.006	0.016	0.585
CL	73	0.016	0.023	0.000	0.002	0.008	0.017	0.112
LLM	398	0.079	0.157	0.000	0.011	0.033	0.082	1.87
T2IM	92.0	0.028	0.043	0.000	0.004	0.014	0.039	0.344
GNN	148.0	0.016	0.029	0.000	0.000	0.006	0.016	0.191

Table 13: Summary statistics of citation density of security papers on six security topics.

Security Topic	Count	Mean	STD	Min	25%	50%	75%	Max
AE	189	0.026	0.063	0.000	0.001	0.006	0.022	0.522
DP	842	0.021	0.044	0.000	0.001	0.008	0.020	0.490
MS	109	0.018	0.036	0.000	0.000	0.006	0.018	0.271
MIA	190	0.023	0.052	0.000	0.002	0.007	0.022	0.585
JB	201	0.112	0.203	0.000	0.023	0.051	0.112	1.872
PI	38	0.070	0.107	0.000	0.009	0.026	0.089	0.433

Table 14: Summary statistics of citation density of security papers with different author counts.

Author	Count	Mean	STD	Min	25%	50%	75%	Max
1	19	0.008	0.014	0.000	0.000	0.002	0.010	0.044
2	153	0.016	0.038	0.000	0.000	0.004	0.013	0.250
3	251	0.028	0.094	0.000	0.001	0.007	0.021	1.192
4	327	0.038	0.125	0.000	0.003	0.009	0.025	1.872
5	267	0.026	0.051	0.000	0.002	0.009	0.026	0.524
6	222	0.040	0.092	0.000	0.003	0.013	0.039	0.916
7	132	0.037	0.090	0.000	0.005	0.014	0.035	0.896
8	55	0.052	0.072	0.000	0.007	0.029	0.071	0.379
9+	55	0.083	0.131	0.000	0.007	0.034	0.093	0.659

Table 15: Summary statistics of citation density of security papers with top 6 affiliated regions.

Region	Count	Mean	STD	Min	25%	50%	75%	Max
AUS	113	0.036	0.081	0.000	0.004	0.011	0.035	0.659
CA	67	0.021	0.036	0.000	0.002	0.007	0.020	0.193
CN	697	0.028	0.062	0.000	0.001	0.009	0.029	0.659
FR	36	0.016	0.037	0.000	0.000	0.005	0.015	0.206
DE	91	0.062	0.204	0.000	0.006	0.017	0.049	1.872
IN	32	0.025	0.063	0.000	0.000	0.003	0.013	0.271
JPN	44	0.034	0.066	0.000	0.000	0.010	0.031	0.379
SGP	78	0.062	0.114	0.000	0.006	0.023	0.065	0.659
KOR	42	0.017	0.050	0.000	0.002	0.005	0.011	0.317
UK	100	0.057	0.195	0.000	0.005	0.014	0.038	1.872
USA	531	0.059	0.142	0.000	0.005	0.016	0.050	1.872

Table 16: Summary statistics of citation density of security papers with different collaboration patterns.

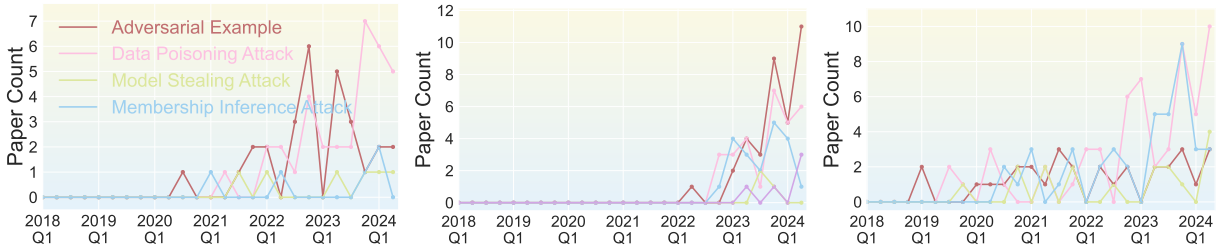
Pattern	Count	Mean	STD	Min	25%	50%	75%	Max
Domestic	486	0.034	0.078	0.000	0.003	0.009	0.028	0.896
Independent	505	0.026	0.081	0.000	0.000	0.007	0.022	1.192
International	484	0.043	0.115	0.000	0.004	0.012	0.039	1.872

Table 17: Summary statistics of citation density of security papers with different publication status.

Status	Count	Mean	STD	Min	25%	50%	75%	Max
Published	1013	0.031	0.077	0.000	0.002	0.009	0.026	1.192
Unpublished	469	0.041	0.118	0.000	0.003	0.011	0.035	1.872

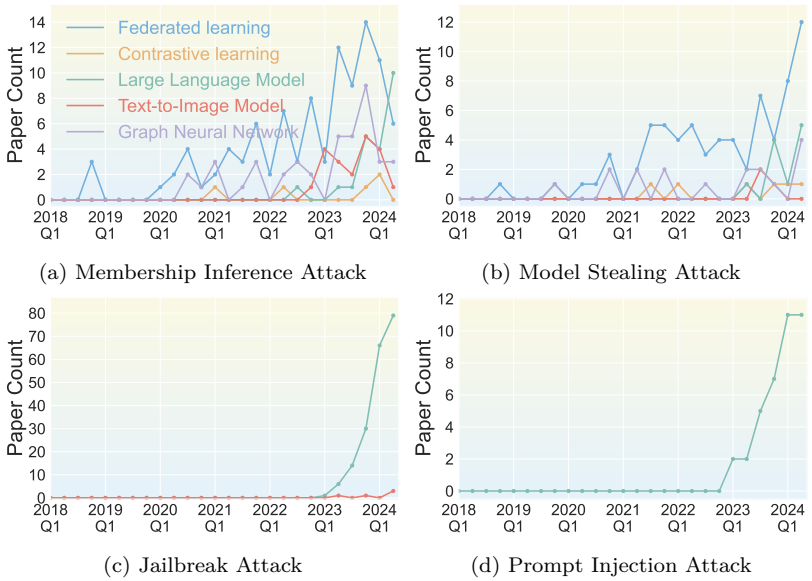
Table 18: Defense lag (in quarters) using target-only entries. Positive values indicate defense trailing attack; negative values indicate defense preceding attack (proactive defense or overlapping topic boundaries). "—" indicates insufficient data.

Topic \ Concept	FL	CL	LLM	T2IM	GNN
AE	1	−4	−4	−1	0
DP	−1	7	5	1	−1
MS	−13	—	−2	—	8
MIA	0	0	3	2	−1
JB	—	—	1	4	—
PI	—	—	1	—	—



(a) Contrastive Learning (b) Text-to-Image Models (c) GNNs

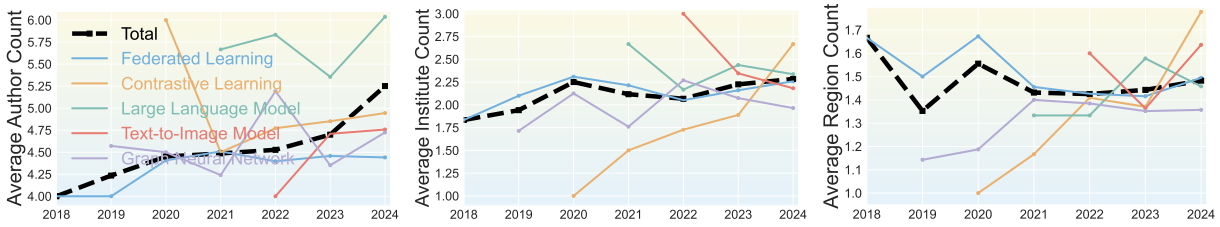
Figure 24: Evolution of the number of security papers for CL, T2IMs, and GNNs. The y-axis represents the number of papers, while the x-axis indicates time, with each data point showing the total number of papers published within a three-month period (a quarter).



(a) Membership Inference Attack (b) Model Stealing Attack

(c) Jailbreak Attack (d) Prompt Injection Attack

Figure 25: Evolution of the number of security papers for model stealing, MIA, jailbreak, and prompt injection. The y-axis represents the number of papers, while the x-axis indicates time, with each data point showing the total number of papers published within a three-month period (a quarter).



(a) Average Author Count (b) Average Institute Count (c) Average Region Count

Figure 26: The Evolution of Average Author Count, Institute Count, and Region Count for Security Papers of 5 ML Concepts

Figure 27: Proportion of three collaboration patterns.