AI Security in the Foundation Model Era: A Comprehensive Survey from a Unified Perspective

Anonymous authors
Paper under double-blind review

Abstract

As machine learning (ML) systems expand in both scale and functionality, the security landscape has become increasingly complex, with a proliferation of attacks and defenses. However, existing studies largely treat these threats in isolation, lacking a coherent framework to expose their shared principles and interdependencies. This fragmented view hinders systematic understanding and limits the design of comprehensive defenses. Crucially, the two foundational assets of ML—**data** and **models**—are no longer independent; vulnerabilities in one directly compromise the other. The absence of a holistic framework leaves open questions about how these bidirectional risks propagate across the ML pipeline. To address this critical gap, we propose a *unified closed-loop threat taxonomy* that explicitly frames model–data interactions along four directional axes. Our framework offers a principled lens for analyzing and defending foundation models. The resulting four classes of security threats represent distinct but interrelated categories of attacks: (1) Data→Data (D→D): including *data decryption attacks, watermark removal attacks, and jailbreak attacks*. (2) Data→Model (D→M): including *poisoning and harmful fine-tuning attacks*; (3) Model→Data (M→D): including *model inversion, membership inference attacks, and training data extraction attacks*; (4) Model→Model (M→M): including *model extraction attacks*. We conduct a systematic review that analyzes the mathematical formulations, attack and defense strategies, and applications across the vision, language, audio, and graph domains. Our unified framework elucidates the underlying connections among these security threats and establishes a foundation for developing scalable, transferable, and cross-modal security strategies—particularly within the landscape of foundation models.

1 Introduction

As The growth of machine learning (ML) has brought about not only more powerful and versatile systems but also an increasingly intricate security landscape. A wide spectrum of threats has emerged—including poisoning, evasion, extraction, and inference attacks—alongside a variety of defensive strategies designed to counter them. While these contributions have advanced the field, they are often examined in isolation, emphasizing case-specific mechanics rather than uncovering the underlying principles that connect them. This siloed treatment fragments our understanding of adversarial behaviors, complicates efforts to reason about their relationships, and hinders the development of defenses that remain effective across diverse attack surfaces. In practice, both researchers and practitioners are left without a coherent framework to navigate the accelerating expansion of ML vulnerabilities.

Underlying this complexity is the fact that the two essential building blocks of ML—**data** and **models**—are deeply interdependent. Compromising data integrity can destabilize or corrupt models, while weaknesses in models can expose private data or propagate errors downstream. Yet, existing surveys rarely capture this mutual influence or explain how risks circulate through the end-to-end ML pipeline. This gap is especially pressing in the context of foundation models, which underpin a wide range of applications and amplify the consequences of security breaches. To address this challenge, we introduce a *unified closed-loop threat taxonomy* that characterizes security dynamics along four directional flows between data (D) and model

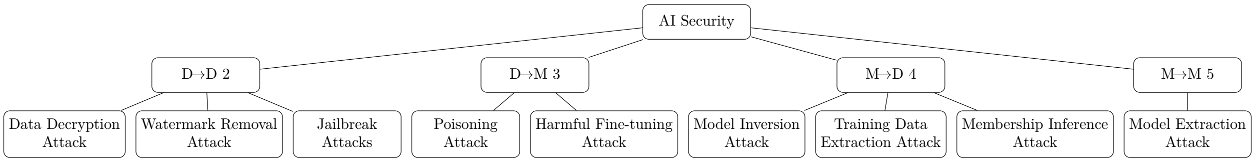

Figure 1: Taxonomy of attacks in AI security

(M): Data → Data (D→D), Data → Model (D→M), Model → Data (M→D), and Model → Model (M→M), as illustrated in Figure 1.

The resulting four classes of security threats represent distinct but interrelated categories of attacks: (1) D→D: This category encompasses attacks that directly manipulate or recover data content. A *data-decryption attack* attempts to recover plaintext information without access to the secret key; a *watermark-removal attack* seeks to eliminate embedded provenance or ownership identifiers; and a *jailbreak attack* constructs adversarial inputs that bypass safety mechanisms, causing the model to disregard policy constraints and carry out unintended behaviors. (2) D→M: *poisoning and harmful fine-tuning attacks* injects malicious samples into training pipelines to induce targeted model misbehavior; (3) M→D: *model inversion, membership inference attacks, and training data extraction attacks* reconstructs sensitive training content from the model's outputs/representations or infers data membership in the training data; (4) M→M: *model extraction attacks* replicates proprietary models via limited model queries. They jointly form a closed-loop of data–model interactions. While each class of threats corresponds to a distinct family of attacks, they are deeply interconnected. For instance, a poisoning attack (D→M) can corrupt the feature–label relationship, thereby weakening the model's robustness to inversion (M→D). Similarly, successful model extraction (M→M) can enable data recovery (M→D) or facilitate adversarial data generation (D→D). These interdependencies are not merely incidental—they form a dynamic chain of influence wherein the compromise of one component (data or model) can recursively propagate vulnerabilities throughout the ML pipeline. Our study offers a comprehensive review of existing research, detailing the mathematical formulations, adversarial and defensive methodologies, and applications spanning visual, linguistic, auditory, and graph-structured data.

As these threats cascade, treating each attack as independent is inadequate. Instead, they motivate a global framework that captures the *full feedback loop* between data and model. This closed-loop perspective is essential for designing robust, scalable, and cross-modal defenses—particularly in the foundation-model era, where data and model boundaries are deeply entangled and increasingly inseparable.

Our contributions can be summarized as:

1. We provide a comprehensive survey that unifies all four data–model attack directions, offering a holistic perspective on the interconnections across threat surfaces.

2. We systematically categorize and analyze representative attacks and defenses within a closed-loop framework, highlighting common principles, differences, and emerging patterns.

3. We identify open challenges and promising research directions, providing guidance for developing more generalizable and resilient defenses against future threats.

Paper Organization. Section 2 introduces D→D threats, including data decryption, watermark removal, and jailbreak attacks and defenses. Section 3 surveys D→M data poisoning and harmful fine-tuning attacks and defenses. Section 4 explores M→D threats such as model inversion, membership inference, and training data extraction attacks and defenses. Section 5 reviews M→M attacks, encompassing data-free, data-based, and architecture cloning attacks and defenses. Finally, Section 6 discusses open challenges and future directions, and Section 7 concludes the paper by summarizing the unified framework and key takeaways.

2 Data→Data (D→D)

2.1 Protection-Bypass Attacks

Protection-bypass attacks Zou et al. (2022; 2025) have become increasingly prevalent in the modern machine learning era, as unauthorized access, misuse, or circumvention of protective mechanisms can lead to severe security and ethical risks. These attacks generally aim to transform protected or original data, or to manipulate model behavior, in order to bypass ownership constraints or safety alignment mechanisms—such as digital encryption, watermarking, or model guardrails. Since classical adversarial attacks are already well established and extensively surveyed in the literature Xiao et al. (2018a;b); Xu et al. (2020); Chakraborty et al. (2021), we do not emphasize them in this survey. Instead, we focus on three representative categories of protection-bypass attacks: *data decryption attacks*, *watermark removal attacks*, and *jailbreak attacks.*

The goal of *data decryption attacks*Laad & Sawant (2021) is to transform encrypted datasets in order to gain unauthorized access or use. Such attacks attempt to recover the original data $\boldsymbol{x}$ from its protected representation $\tilde{\boldsymbol{x}}$ without possessing the secret key.

The goal of *watermark removal attacks*Zhao et al. (2024a) is to transform a watermarked data $\tilde{\boldsymbol{x}}$ (image, or text) into another data $\boldsymbol{x}'$ that preserves task utility while disabling watermark detection or decoding. The attack operates solely through low-level (e.g., pixel- or feature-level) or content-level transformations of the dataset, without accessing the model parameters. The scope includes visible marks (logos, stamps, overlays) and invisible marks embedded in the spatial, frequency, or learned feature domains.

The goal of *jailbreak attacks* Liu et al. (2025a) is to bypass the safety alignment of large foundation models and trick them into producing harmful or restricted outputs. Unlike adversarial attacks that cause simple misclassification, jailbreaks directly override ethical or safe policy constraints. By breaking built-in safeguards, jailbreaks can lead to misinformation, malicious code, or other unauthorized behaviors.

Common Procedures. Despite their differences, these attacks share a common operational structure. Attackers typically optimize a transformation that maps a protected input $\tilde{\boldsymbol{x}}$ to an output $\boldsymbol{x}'$, satisfying two conditions: (i) the ownership constraint is removed (e.g., watermark undetectable, cipher broken, or safety bypassed); (ii) the transformed data retains high visual or semantic fidelity to the original. For *image data*, typical pipelines first localize protected regions (e.g., watermarked or encrypted areas) and then restore or regenerate them. Traditional signal-processing operations such as compression, denoising, and filtering remain strong baselines for disrupting watermark detectors, while modern attacks increasingly use generator-based regeneration or latent-space resampling—e.g., encode–decode pipelines with diffusion models or VAEs—to perform end-to-end rewriting Zhao et al. (2024a). For *text data*, token-level perturbations (insertion, deletion, substitution), paraphrasing, back-translation, or guided rewriting can weaken watermark statistical signals and evade detection Piet et al. (2025). These rewriting-based transformations, akin to prompt manipulations in *jailbreak attacks*, leverage semantic-preserving reformulations to bypass ownership or alignment constraints while maintaining fluency and meaning.

2.2 Mathematical Formalization

General Objective: Let $\boldsymbol{x} \in \mathcal{X}$ denote a clean data sample. A protection mechanism P transforms $\boldsymbol{x}$ into a protected form $\tilde{\boldsymbol{x}}$ by applying ownership or safety constraints such as encryption, watermarking:

$$\tilde{\boldsymbol{x}} = P(\boldsymbol{x}; \boldsymbol{k}),$$

where $\boldsymbol{k}$ denotes a secret key (for encryption and authorized decryption), a watermark identifier, or may be null (so P reduces to the identity). The adversary aims to construct an operator $U \in \mathcal{U}$ that maps $\tilde{\boldsymbol{x}}$ to a surrogate $\boldsymbol{x}' = U(\tilde{\boldsymbol{x}})$ that (i) recovers the original content without the decryption key, (ii) removes or invalidates embedded ownership marks, or (iii) manipulates aligned models into producing harmful or policy-violating outputs, while preserving the utility or perceptual fidelity of the original sample.

The attack objective can be written as

$$\begin{aligned} \max_{U \in \mathcal{U}} \quad & O[V(U(\tilde{\boldsymbol{x}})) = 1] \\ \text{s.t.} \quad & S(\boldsymbol{x}', \boldsymbol{x}) \;\geq\; \tau, \qquad C(U) \;\leq\; B. \end{aligned} \tag{1}$$

Here, V denotes a verification function that outputs 1 when the attack is successful and 0 otherwise. The probability operator O is taken over the data distribution $\boldsymbol{x}$, representing the likelihood that verification still succeeds after the adversarial transformation. Specifically: for *(i) data decryption attacks*, V verifies whether unauthorized decryption succeed. A successful attack means that the original data can be recovered without the secret key $\boldsymbol{k}$. For *(ii) watermark removal attacks*, V verifies whether the embedded watermark remains detectable. A successful attack means that the watermark becomes undetectable. For *(iii) jailbreak attacks*, V evaluates whether the transformed input can successfully bypass the model's safety or policy enforcement. A successful attack means that the model's response violates safety alignment. The similarity metric $S(\cdot, \cdot)$ ensures perceptual or utility preservation relative to the original sample $\boldsymbol{x}$; τ denotes a quality threshold, and C with budget B represents the computational or query cost constraint.

2.2.1 Special Case I

Data Decryption AttacksLaad & Sawant (2021). When the protection mechanism is encryption, $U = D$ is an decryption function:

$$\boldsymbol{x}' = D(\tilde{\boldsymbol{x}}),$$

and V verifies whether the recovered sample $\boldsymbol{x}'$ successfully reconstructs the original data, implying that the encryption has been effectively reversed. The attack seeks to approximate the decryption process without access to the secret key, producing $\boldsymbol{x}' \approx \boldsymbol{x}$.

2.2.2 Special Case II

Watermark Removal AttacksZhao et al. (2024a). When the protection mechanism is watermarking, $U = A$ is a watermark removal attacks, i.e.,

$$\boldsymbol{x}' = A(\tilde{\boldsymbol{x}}),$$

and V is a detection or decoding function D that verifies the presence of watermark. A successful attack means that the embedded watermark is no longer detectable. At the same time, $\boldsymbol{x}'$ should preserve utility for downstream tasks.

2.2.3 Special Case III

Jailbreak Attacks Liu et al. (2025a). When the protection mechanism is safety alignment, the attacker aims to construct a transformation $U = J$ that modifies the input prompt $\tilde{\boldsymbol{x}}$ into a new query $\boldsymbol{x}' = J(\tilde{\boldsymbol{x}})$ which bypasses alignment or moderation constraints. Here, V denotes a safety verification function that outputs 1 if the model's output satisfies alignment rules (e.g., refuses to produce harmful content). Here, $\boldsymbol{x}'$ must remain a valid query and preserve coherence so that the model generates meaningful but unauthorized outputs.

Attacker Knowledge. We use the term *detector* to denote the algorithm used to identify ownership signals, and the *key* refers to the hidden parameter or seed controlling embedding and decoding. The strength of an attack depends on the attacker's knowledge of the protection mechanism Kirchenbauer et al. (2023). In a *black-box* setting, the adversary can only query a detector and observe binary outputs, with no access to the secret key or model internals. In a *gray-box* setting, partial information such as the architecture or general watermarking algorithm is available, but the secret key remains unknown. In a *white-box* setting, the attacker has full access to the detection system and can directly manipulate internal parameters to design optimized attacks.

Attack effectiveness is typically evaluated along two axes. First, *post-attack verifiability* measures the remaining strength of ownership or alignment constraints after the attack, i.e., the probability that the verification function V still outputs 1 following transformation by U. Lower verifiability indicates more successful removal or circumvention of the protection mechanism. For example, in watermarking, this corresponds to post-attack detectability or decoding accuracy. Second, *utility preservation* assesses how much of the original data's perceptual, semantic, or functional quality is retained for downstream tasks. A successful attack achieves low verifiability while maintaining high utility, highlighting the inherent trade-off between stealth and fidelity.

2.3 Taxonomy and Techniques of Protection-Bypass Attacks

We broadly divide the D→D attacks into three major categories: *1) Data Decryption Attacks*, which attempt to recover original content from encrypted data, *2) Watermark Removal Attacks*, which aim to erase ownership signals embedded in data, and *3) Jailbreak Attacks*, which bypass alignment or safety rules in foundation models to force them to produce restricted or harmful content.

2.3.1 Data Decryption Attacks

These attacks aim to recover or approximate original data from encrypted data without possessing the secret key. We categorize them into four main families according to strategy and adversarial assumptions: *(a) key-recovery and cryptanalysis*, exploiting brute-force, weak diffusion, or statistical dependencies in chaos-based cryptosystems Guan et al. (2005); Fridrich (1998); *(b) ciphertext-only and statistical reconstruction*, where attackers infer data distributions or visual content directly from ciphertext features—such as in GAN- or feature-based ciphertext-only attacks Sirichotedumrong & Kiya (2020); *(c) generative-model and learning-based regeneration*, leveraging pretrained generative priors (e.g., GANs or diffusion models) to reconstruct visually plausible plaintexts MaungMaung & Kiya (2023); and *(d) side-channel and leakage-based attacks*, where partial computational leakage (e.g., timing or memory access) undermines key secrecy or enables partial recovery Benhamouda et al. (2018).

2.3.2 Watermark Removal Attacks

Watermark removal seeks to erase ownership signals while maintaining perceptual or semantic fidelity. Approaches can be roughly divided by operating space: *(a) pixel- and signal-space distortions* apply JPEG compression, filtering/denoising, noise injection, resampling, rotation, scaling, cropping, or affine transforms as classical baselines Wan et al. (2022); Begum & Uddin (2020); Mousavi et al. (2014); *(b) Mask-guided detection and inpainting* methods adopt a two-stage "localize–then–restore" paradigm based on decomposition or refinement networks. In the first stage, the attacker explicitly detects or localizes the watermark region by generating a mask that estimates the opacity, color, or spatial extent of the mark. In the second stage, the masked area is filled in by an inpainting or restoration network to recover the original image content hidden beneath the watermark Liu et al. (2021); Liang et al. (2021); Zhao et al. (2022); Niu et al. (2023). *(c) Generator-based regeneration*, where architectures such as convolutional, transformer, or disentangled networks Li et al. (2021a); Sun et al. (2023) learn a single-branch end-to-end mapping that directly translates a watermarked input into a clean output without producing any intermediate mask, and *(d) latent-space resynthesis*, where diffusion or VAE models remove watermarks in the representation space Su & Zhang (2025); Zhao et al. (2024a); Liu et al., this process effectively bypasses localized pixel- level traces of the watermark. For text, *(e) editing- and paraphrasing-based attacks*, which weaken watermark signals by applying semantics-preserving transformations such as word substitution, paraphrasing, or style rewriting that alter token statistics or sentence structure to evade detection Yang et al. (2025b); Kirchenbauer et al. (2023); Liu et al. (2024a), while *(f) Model-driven approaches* infer or neutralize watermarking rules directly through surrogate modeling or decoding-time neutralization Pan et al. (2025). Recent surveys highlight a paradigm shift from heuristic distortions toward learning-based regeneration methods that balance watermark removal effectiveness with perceptual or semantic fidelity Su & Zhang (2025); Liu et al. (2024a); Wan et al. (2022).

2.3.3 Jailbreak Attacks

Jailbreaks bypass alignment or safety mechanisms to force models to output restricted or harmful content that would normally be blocked, posing a major threat to LLMs and multimodal models. They can be characterized by: *(a) attacker access*, distinguishing white-box gradient-guided suffix generation from black-box prompt rewriting and role-play prompting Geisler et al.; *(b) attack timing*, including training-stage backdoor injection and inference-stage adversarial prompting Chao et al. (2024); *(c) prompt manipulation strategies*, such as template completion, scenario nesting, cipher-based rewriting, or genetic optimization Chao et al. (2024); and *(d) system-level extensions*, where multimodal or agentic systems are compromised through adversarial cross-modal cues or external tools/APIs/ memory manipulation for autonomous agents Xu et al. (2024b); Liu et al. (2025a). Subsequent studies have expanded jailbreak analyses to multimodal systems. *Dong et al.* Dong et al. (2023) systematically evaluate adversarial image attacks against Google's Bard, revealing that small, imperceptible perturbations can bypass both face- and toxicity-detection modules, thereby exposing critical vulnerabilities in visual-language alignment and safety filtering of commercial MLLMs. More recently, research has shown that such system-level vulnerabilities also extend to agentic architectures. For example, Li et al. Li et al. (2025a) demonstrate that even commercial LLM-based web and scientific agents are vulnerable to trivial yet dangerous prompt-injection and redirection attacks. These attacks demonstrate how behavioral safeguards can be circumvented even without modifying model parameters.

2.4 Defensive Techniques

Defenses against data→data attacks share a common goal: preserve data ownership and safe use under strong, adaptive adversaries. We organize defenses into three families aligned with the attacks reviewed above: 1) against *data decryption*, strengthening ciphers and enabling privacy-preserving computation; 2) against *watermark removal*, reinforcing embedding, detection, tamper localization/recovery, and proactive protection; and 3) against *jailbreaks*, strengthening prompts, models, and deployed systems to resist alignment bypass and safety circumvention Mao et al. (2025); Liu et al. (2025a).

2.4.1 Defenses against Data Decryption Attacks

Existing approaches fall into three broad classes: *(a) Chaos–neural hybrids*, which enlarge key space and resist statistical or differential attacks by combining chaotic maps with neural networks Lakshmi et al. (2021); *(b) Autoencoder-/GAN-based encryption*, which hide data via Cycle-GAN-based transformations that map images into hard-to-invert hidden domains Ding et al. (2020); *(c) Privacy-preserving learning on encrypted data* using homomorphic encryption and secure multi-party computation, enabling distributed model training Tang et al. (2019) and encrypted-domain inference Bost et al. (2014) without ever exposing plaintext. These directions complement system hygiene against *side-channel leakage* and integrity risks Benhamouda et al. (2018); Manikandan & Masilamani (2018).

2.4.2 Defenses against Watermark Removal Attacks

We group defenses against Watermark Removal Attacks into four dimensions: *(a) Robust embedding* strengthens both visible and invisible watermark signals by integrating multiple embedding strategies. For visible marks, recent works employ multi-level alpha blending, adaptive texture-aware placement, and randomized geometric positioning to make watermark removal leave noticeable artifacts Dekel et al. (2017). For invisible marks, robustness is achieved through transform-domain embedding (e.g., discrete cosine transform modification of frequency coefficients Barni et al. (1998)), spread-spectrum coding with redundancy across channels Cox et al. (1997), and synchronization patterns that maintain watermark alignment under rotation, scaling, or cropping Lin & Chang (1997). Together, these designs form the foundation of modern invisible and visible watermark protection Wan et al. (2022). *(b) Robust Detection and Verification* enhance the reliability of watermark verification through stronger statistical testing, improved scoring, and resilient encoding. For text and code watermarking, recent studies propose *finite-sample hypothesis tests* that avoid Gaussian approximations and enable more accurate detection under limited data conditions Liu et al. (2024a); Yang et al. (2025b). Meanwhile, the work Golowich & Moitra (2024) analyzes the vulnerability of pseudo-random indexing watermarks under editing attacks, showing that robustness provably degrades with increasing edit distance. Together, these strategies form a layered defense, coupling operational safeguards with theoretical robustness for trustworthy watermark verification; *(c) Manipulation detection and content recovery*, which co-embed localization and self-recovery signals to make tampering both detectable and, when possible, reversible for image and document integrity, respectively Ying et al. (2023); Cui et al. (2024); and *(d) Proactive* *data protection (Unlearnable Examples)*, which pre-perturbs data so that unauthorized training fails while preserving data utility, encompassing robust/stable Liu et al. (2024c), semantic or feature-space perturbation defenses Meng et al. (2024), transferable, model-free, and surrogate-free Sadasivan et al. (2023), and cross-modal extensions with theoretical motivation Jiang et al. (2024).

2.4.3 Defenses against Jailbreak Attacks

We organize jailbreak attack defenses into three main categories based on the protection level: *(a) Prompt-level* screening and rewriting—detecting risky or injected inputs via fine-tuned classifiers and heuristic filters, and sanitizing or rewriting them before the model processes the prompt Jacob et al. (2024); *(b) Model-level* alignment and steering—reinforcement learning from human feedback (RLHF) Ouyang et al. (2022) and related safety fine-tuning enhance model alignment and refusal behavior, while decoding-time constraints and internal-signal detectors further mitigate unsafe generations Ouyang et al. (2022); and *(c) System-level* guardrails—benchmark-driven guard models, multi-stage filters for multimodal LLMs, tool/memory governance for agents, and continuous runtime monitoring at deployment Huang et al. (2024e); Chao et al. (2024). Recent advances expand this line of work: *AIR-BENCH 2024* Zeng et al. introduces a regulation-aligned auditing framework that evaluates model refusal behaviors and compliance across real-world risk categories; *T2VSafetyBench* Miao et al. (2024) extends evaluation to text-to-video generative models, revealing multimodal jailbreak vulnerabilities; and *AEGIS-LLM* Cai et al. demonstrates that incorporating auxiliary agent roles and leveraging automated prompt optimization can enhance system robustness without compromising task utility. In practice, these defenses must account for diverse attack settings—including both white- and black-box access, and attacks occurring at training or inference time—such as suffix optimization, backdoor injection, role-play or cipher-based prompt rewriting, and evolutionary prompt search Liu et al. (2025a).

3 Data→Model (D→M)

Modern machine learning systems rely on large and diverse datasets to train foundation models. Because training pipelines may include data from untrusted or unverified sources, they present significant security and privacy risks. *D→M attacks* exploit this vulnerability by manipulating training or fine-tuning data to implant malicious behaviors or bypass safety alignment. Unlike *D→D* attacks that modify the data itself, D→M attacks corrupt the learning process, causing models to internalize unintended objectives and degrade in reliability. We focus on two representative families of D→M attacks: *data poisoning* and *harmful fine-tuning.*

The goal of *data poisoning attacks* is to corrupt the training data so that the model learns attacker-specified behaviors or incorrect objectives. Unlike test-time adversarial examples that perturb inputs after deployment, poisoning intervenes during model training or pre-training. Attackers modify samples, labels, or data flows so that empirical risk minimization optimizes a malicious objective, causing the resulting parameters to embed hidden biases or backdoors Chen et al. (2017); Tian et al. (2022); Cinà et al. (2023). Such attacks can degrade model accuracy, trigger targeted misclassification, or implant covert functionality that persists even after alignment fine-tuning. They have been demonstrated across domains—from image and graph learning to LLM instruction tuning Geiping et al. (2021) —highlighting the vulnerability of model optimization in distributed and federated training settings.

Fine-tuning has become the standard mechanism for adapting foundation models to specific tasks. However, granting public or API-level access to fine-tuning pipelines introduces a new threat. The goal of *harmful fine-tuning attacks* is to manipulate the fine-tuning stage of foundation models to bypass safety alignment and induce undesired behaviors. Attackers upload small but malicious datasets—sometimes containing seemingly benign samples—to bias the model toward unsafe or target-specific responses. Unlike full retraining, these attacks require minimal data and computation yet can cause significant shifts in model behavior by exploiting the sensitivity of alignment layers. Recent studies Qi et al. (2024) demonstrate that even minimal fine-tuning—using only a handful of carefully crafted examples—can drastically undermine alignment, suppress refusal behaviors, and enable the generation of restricted content, highlighting the fragility of current safety mechanisms.

3.1 Mathematical Formalization

A unified view of D→M attacks is that the adversary uses a model $f_{\boldsymbol{\theta}}$, a benign dataset $\mathcal{D}_{\text{clean}}$, together with a generation strategy $\mathrm{G}(f_{\boldsymbol{\theta}}, \mathcal{D}_{\text{clean}})$ to produce a small set of malicious training or fine-tuning samples $\mathcal{D}_{\text{adv}}$, which are then injected into the benign dataset $\mathcal{D}_{\text{clean}}$ back. The goal is that the later resulting learning process yields parameters $\boldsymbol{\theta}$ that satisfy an attacker-specified objective while remaining stealthy on the model's normal tasks. This interaction can be formulated as a bilevel optimization:

$$\begin{aligned} &\mathcal{D}^*_{\text{adv}} \;=\; G(f_{\boldsymbol{\theta}^*}, \mathcal{D}_{\text{adv}}) \\ &\text{s.t.} \quad \boldsymbol{\theta}^*(\mathcal{D}_{\text{clean}} \cup \mathcal{D}_{\text{adv}}) = \arg\min_{\boldsymbol{\theta}} \; \mathcal{L}_{\text{train}}(\boldsymbol{\theta};\; \mathcal{D}_{\text{clean}} \cup \mathcal{D}_{\text{adv}}) \\ &\qquad\; S(\mathcal{D}_{\text{adv}}, \mathcal{D}_{\text{clean}}) \;\geq\; \tau. \end{aligned}$$

$\mathcal{L}_{\text{train}}$ is the training (or fine-tuning) loss minimized on the dataset $(\mathcal{D}_{\text{clean}} \cup \mathcal{D}_{\text{adv}})$; it encodes the attacker's objective (e.g., increasing the rate of unsafe responses, forcing misclassification on triggered inputs, or reducing overall utility). The constraint S is a similarity metric that expresses stealth (e.g., small cardinality, bounded perturbation, distributional similarity), τ is a quality threshold.

3.1.1 Special Case I

The goal of **data poisoning attacks** Tian et al. (2022) is to add or modify training samples so that the learned model parameters $\boldsymbol{\theta}^*$ produce attacker-specified failures (untargeted degradation) or targeted misbehaviour (backdoors). Concretely, the attacker constructs a poisoned dataset $\mathcal{D}_{\text{poison}}$ — generated via label manipulation, optimization-based perturbations, or related techniques — to optimize an adversarial objective $\mathcal{L}_{\text{adv}}$. Using the notation above, the poisoning optimization is:

$$\begin{aligned} \mathcal{D}^*_{\text{adv}} = \{(\boldsymbol{x}'_j, y'_j)\}_{j=1}^m = G(f_{\boldsymbol{\theta}^*}, \mathcal{D}_{\text{adv}}) \\ G(f_{\boldsymbol{\theta}^*}, \mathcal{D}_{\text{adv}}) := \arg\min_{(\boldsymbol{x}', y')} \mathcal{L}_{\text{adv}}\big(\boldsymbol{\theta}^*(\boldsymbol{x}', y'); \mathcal{D}_{\text{target}}\big) \\ \text{s.t.} \quad \boldsymbol{\theta}^*(\mathcal{D}_{\text{poison}}) = \arg\min_{\boldsymbol{\theta}} \; \mathcal{L}_{\text{train}}(\boldsymbol{\theta}; \mathcal{D}_{\text{poison}}), \\ \mathcal{D}_{\text{poison}} = \mathcal{D}_{\text{clean}} \cup \{(\boldsymbol{x}'_j, y'_j)\}_{j=1}^m, \|\boldsymbol{x}'_j - \boldsymbol{x}_j\|_p \leq \epsilon. \end{aligned}$$

where := denotes definition, $(\boldsymbol{x}_j, y_j)$ denotes a clean training sample and $(\boldsymbol{x}'_j, y'_j)$ its poisoned counterpart, $y'_j \neq y_j$. $\mathcal{D}_{\text{target}}$ is a fixed set of target examples used to evaluate or trigger the adversarial objective (not an optimization variable), m is the number of injected poisoning points, and ϵ controls the maximum perturbation magnitude under the L_p norm.

Common instances: - *Untargeted poisoning*: maximize the global validation loss or reduce overall accuracy on a held-out validation set. - *Targeted/backdoor poisoning*: choose $\mathcal{D}_{\text{target}}$ to encode specific inputs $\boldsymbol{x}^*$ (with trigger) that should map to y'_j, y'_j is the target label, y_j is the true label, $y'_j \neq y_j$.

3.1.2 Special Case II

The goal of **harmful fine-tuning attacks** Halawi et al. is to compromise an aligned or pre-trained model by providing a small fine-tuning dataset $\mathcal{D}_{\text{adv}}$ (via API or shared service) so that the adapted parameters $\boldsymbol{\theta}_{\text{ft}}$ exhibit unsafe or biased behavior while retaining nearly unchanged performance on benign tasks, here G constructs $\mathcal{D}_{\text{adv}}$ by following a set of predefined rules and procedures. The attack can be formulated as:

$$\begin{aligned} \mathcal{D}^*_{\text{adv}} = \{(\boldsymbol{x}_j, \boldsymbol{y}'_j)\}_{j=1}^m = G(f_{\boldsymbol{\theta}^*}, \mathcal{D}_{\text{adv}}) := W(\{(\boldsymbol{x}_j, \boldsymbol{y}_j)\}_{j=1}^m) \\ \text{s.t.} \quad \boldsymbol{\theta}^* = \boldsymbol{\theta}_0, \end{aligned}$$

Here W denotes the attacker's data-modification rules mapping a clean example $(\boldsymbol{x}, \boldsymbol{y})$ to a harmful example $(\boldsymbol{x}, \boldsymbol{y}')$. $\boldsymbol{\theta}_0$ are the original parameters. Typical attacker targets include removing refusal behavior, inserting triggers, or shifting outputs toward biased content. Such attacks can exploit parameter-efficient fine-tuning modules—e.g., LoRA adapters or prompt layers Gao et al. (2024a)—often requiring only a few crafted examples and can be hard to detect.

3.2 Taxonomy and Techniques of Data→Model Attacks

In practice, data→model attacks fall into two broad families: *data poisoning* refers to the manipulation of training data such that standard learning procedures inadvertently optimize an adversarial objective; and *harmful fine-tuning* uses small, carefully curated datasets to degrade a model's safety alignment or to implant malicious behaviors. In the following, we highlight the key taxonomic axes, common operational mechanisms, and representative works.

3.2.1 Data Poisoning Attacks

Poisoning attacks manipulate training data to implant malicious behaviors or degrade model reliability. They can be systematically characterized along four orthogonal dimensions—*adversarial goals*, *label visibility*, *attacker knowledge*, and *training paradigm*—as discussed in recent comprehensive surveys Tian et al. (2022); Nguyen et al. (2024b); Rodríguez-Barroso et al. (2023). Below, we summarize each axis and its representative techniques. *(a) Adversarial Goals.* Poisoning objectives are commonly divided into *availability attacks*, which reduce overall model utility (e.g., accuracy or calibration), and *integrity attacks*, which implant targeted behaviors such as backdoors or hidden triggers that activate only under specific conditions Gu et al. (2017). Recent work further identifies more subtle objectives, such as degrading model confidence or calibration without label manipulation Chaalan et al. (2024). *(b) Label Visibility.* Depending on whether the attacker manipulates labels, poisoning can be either *dirty-label* or *clean-label*. Dirty-label attacks explicitly flip or corrupt training labels, whereas clean-label attacks preserve ground-truth labels but perturb inputs to mislead the learned decision boundary—typically by generating adversarial-like examples that induce feature collisions Shafahi et al. (2018). *(c) Attacker Knowledge.* The strength and strategy of poisoning attacks vary with the attacker's access level. In *white-box* settings, full access to model parameters and gradients enables optimization-based attacks, typically formulated as bilevel optimization and often approximated through influence functions Koh et al. (2022). *Gray-box* attackers possess partial information—such as model architecture or aggregation rules in federated learning—and adapt their poisons accordingly Rodríguez-Barroso et al. (2023). In contrast, *black-box* attackers lack internal access but can still exploit transferability from surrogate models Zhu et al. (2019), demonstrating that effective poisoning remains feasible even without visibility into gradients or training data Chen et al. (2023); Liu & Lai (2021). *(d) Training Paradigm.* Poisoning manifests differently across learning paradigms. In conventional centralized training, attackers can directly manipulate datasets by injecting crafted samples or flipping labels Ramirez et al. (2022). In federated learning, poisoning Xie et al. arises through malicious client updates or model replacement, where even a single compromised participant can bias the global aggregation and degrade overall model integrity Nguyen et al. (2024b). Similar paradigm-specific threats extend beyond vision to other modalities—graphs (node or edge injection) Cinà et al. (2023); Tao et al. (2021), time series (temporal or phase triggers) Lin et al. (2024), and language (rare-token, syntactic, or instruction-pattern triggers that persist through fine-tuning) Kurita et al. (2020). At the scale of foundation models, even tiny poisoning ratios can survive model-level safety alignment and propagate through subsequent updates Bowen et al. (2025); Carlini et al. (2024a), highlighting the need for domain-aware validation pipelines in high-stakes applications such as biomedical large language models Alber et al. (2025).

3.2.2 Harmful Fine-tuning Attacks

Harmful fine-tuning manipulates the adaptation stage of aligned models using small curated datasets, leading to unsafe or biased behaviours while retaining benign utility. Existing studies reveal several recurring patterns, summarized below. *(a) Malicious Data Poisoning.* Attackers deliberately insert adversarial prompt–response pairs into the fine-tuning dataset to overwrite safety alignment. As demonstrated in work Yi et al. (2024), applying *reverse alignment*—either via Reverse Supervised Fine-Tuning or Reverse Preference Optimization—can fine-tune safety-aligned open-access LLMs to undermine refusal behaviors and weaken built-in safeguards. *(b) Benign-Data–Induced Misalignment.* Even datasets without explicit harmful content can degrade safety: outlier yet non-toxic examples, distributional biases, or latent correlations in seemingly benign corpora may erode alignment to a degree comparable with adversarial fine-tuning He et al.. These results highlight that alignment failure can arise naturally from poor data curation rather than deliberate attacks. *(c) Parameter-Efficient and Model-Specific Pathways.* Attackers exploit architectural properties and lightweight adaptation mechanisms to inject harmful logic efficiently. Single-edit or adapter-based backdoors, targeted knowledge editing, and PEFT-as-attack demonstrate how small parameter updates can trigger disproportionate behavioural changes Chen et al.. *(d) API Exploitation and Service Abuse.* In hosted environments, attackers exploit fine-tuning APIs to upload covert malicious data. Recent work Halawi et al. shows that pointwise-undetectable datasets, mixed compliance–refusal fine-tuning objectives, and constrained black-box jailbreaks can degrade safety even without access to model weights. Related findings indicate that open fine-tuning interfaces remain an inherent security risk for both commercial and open-source LLMs. Beyond these categories, recent studies reveal that harmful fine-tuning can also emerge in more subtle forms. Emergent misalignment may appear during narrow-domain or agentic fine-tuning, where self-updating models gradually drift toward unsafe policies Hahm et al. (2025); Wallace et al. (2025); Shao et al. (2025b). Together, these observations show that even small, seemingly benign datasets—whether maliciously designed or inadvertently mis-specified—can reliably steer large foundation models toward unsafe or biased behaviours, underscoring the fragility of current alignment processes.

3.3 Defensive Mechanisms

3.3.1 Defense against Data Poisoning

Defenses against data poisoning aim to preserve model integrity despite the presence of malicious or corrupted data. They can be broadly organized by where the intervention occurs: *during training* or *at inference.* Across all settings, effective defense combines anomaly detection, robust optimization, and trust-aware data filtering to limit the adversary's impact. *(a) Training-Time Defenses.* Training-stage defenses focus on identifying or neutralizing poisoned data before or during optimization. Data-level regularization via *augmentation* (e.g., Mixup, CutMix) helps dilute backdoor triggers and reduces memorization of malicious patterns Borgnia et al. (2021). At the loss-function level, *robust learning objectives* derived from noisy-label and meta-learning literature—such as ITLM (Iterative Trimmed Loss Minimization) Shen & Sanghavi (2019), GCE Zhang & Sabuncu (2018), reweighting-based methods Ren et al. (2018), Co-teaching Han et al. (2018), and MentorNet Jiang et al. (2018)—limit the influence of high-loss or inconsistent samples. Feature-space filtering methods like *De-Pois* Chen et al. (2021a) further cluster examples by representation consistency to remove those with mismatched feature–label semantics. A framework *Neural Attention Distillation (NAD)* Li et al. was proposed to use a finetuned teacher network to guide a backdoored student model via attention alignment on a small clean subset. Building on this line, *Li et al.* Li et al. (2021b) introduced *Anti-Backdoor Learning (ABL)*, a training-time paradigm that aims to train clean models directly on poisoned data. However, most training-time defenses remain largely ineffective against *clean-label* or *stealthy backdoor* attacks, in which poisoned samples appear statistically benign and evade standard anomaly detectors Koh et al. (2022); Geiping et al. (2021).

(b) Inference-Time Defenses. Inference-time defenses operate after model deployment and aim to detect or mitigate triggered behavior at test time. A first line of defense is *anomaly detection*, which removes outlier samples using statistical or representation-based criteria such as clustering, spectral signatures, or isolation forests Cinà et al. (2023). *Uncertainty-based filtering* evaluates prediction entropy or stability under perturbations: low-entropy or invariant outputs often reveal the presence of triggers. Typical examples include STRIP for vision Gao et al. (2019), which uses entropy-based uncertainty signals to identify poisoned inputs. Other work Liu et al. (2023) tests robustness under input corruptions (noise, blur, occlusion), exploiting the observation that poisoned examples remain unusually stable under such changes. A complementary direction is *knowledge-guided validation*, which cross-checks model predictions against external knowledge sources, such as biomedical knowledge graphs, to flag implausible outputs Alber et al. (2025).

3.3.2 Defense against Harmful Fine-tuning

Defenses against harmful fine-tuning aim to preserve model safety and alignment even when adversaries attempt to retrain or adapt models with malicious or misleading data. Existing methods can be broadly grouped into three complementary directions. *(a) Alignment-Stage Immunization.* These approaches fortify models during the initial alignment phase to make them intrinsically resistant to future harmful fine-tuning. Representative methods enhance weight stability, representation invariance, or regularization against adversarial updates—such as perturbation-aware and layer-wise robustness training (*Vaccine* Huang et al. (2024d), *T-Vaccine* Liu et al. (2025b)), loss-based regularization and proximal optimization (*Booster* Huang et al. (2024c), *LISA* Huang et al. (2024b)), and representation-level noise injection (*RepNoise* Rosati et al. (2024b)). Formal analyses further define theoretical *immunization conditions*—including resistance, stability, and generalization—that guide preventive alignment strategies Rosati et al. (2024a). *(b) In-Training Safeguards.* These defenses are applied at fine-tuning time to monitor and mitigate malicious model updates in real time. *Self-Degraded Defense (SDD)* Chen et al. (2025b) pre-emptively trains models by pairing harmful prompts with benign, high-quality responses, thereby reducing the model's sensitivity to malicious data while preserving its normal capabilities. For parameter-efficient fine-tuning (PEFT), *PEFTGuard* Sun et al. (2025) detects backdoored adapters by directly transforming and classifying their weight tensors (e.g., LoRA) with a parameter-only meta-classifier, and identifying backdoor-specific patterns with near-perfect accuracy across tasks. Bayesian data scheduler Hu et al. (2025) incorporates probabilistic safety control by assigning weights to samples according to their posterior safety attributes during fine-tuning, effectively suppressing the influence of unsafe or malicious data on model adaptation. *(c) Post-Tuning Repair.* When harmful fine-tuning has already occurred, post-hoc repair methods attempt to recover safety without retraining from scratch. *Antidote* Huang et al. (2024a) prunes harmful parameters identified via importance scoring, effectively restoring alignment with minimal loss in utility. Such an approach treats fine-tuning as reversible damage, focusing on repairing rather than preventing misalignment.

4 Model→Data (M→D)

Attacks in the Model→Data direction aim to infer the information that a trained model implicitly encodes about its training data. Rather than stealing parameters or manipulating the model externally, these attacks exploit what the model *memorizes*—its ability to reveal, reconstruct, or statistically expose private training data set samples. Such leakage undermines data confidentiality and consent, as even deployed or API-restricted models may inadvertently disclose sensitive content through their outputs, embeddings, or confidence patterns. Within this category, we highlight three representative families of privacy-violating attacks: *model inversion attacks*, *membership inference attacks*, and *training data extraction attacks*.

The goal of *model inversion attacks* Fredrikson et al. (2015); Yang et al. (2025a); Dibbo (2023) is to reconstruct sensitive information about the training data directly from a trained model. By exploiting confidence scores, embeddings, or gradients, an adversary can approximate original data features or even recover realistic samples such as faces or text segments. Early studies assumed white-box access, but recent work shows that inversion can succeed in black-box APIs by leveraging systematic output patterns. These attacks reveal how models encode detailed traces of their training data even without direct access to the dataset itself.

The goal of *membership inference attacks* is to determine whether a specific data instance—or an entire subset—was used in a model's training process. By probing the model and analyzing statistical differences between "seen" and "unseen" samples—often manifested in confidence scores, loss values, or hidden representations—attackers can infer a data's participation in sensitive datasets such as medical or personal records. Recent research extends these attacks beyond conventional classifiers to LLMs and diffusion models, where memorization and overfitting amplify data membership signals.

The goal of *training data extraction attacks* is to induce a model to directly reproduce fragments of its original training data, rather than merely inferring or reconstructing them statistically. Through carefully crafted prompts, triggers, or fine-tuning procedures, adversaries can compel LLMs or diffusion models to regenerate exact text passages, images, or identifiers memorized during pre-training. Unlike inversion or membership inference, these attacks cause the model to emit verbatim training content, posing severe risks to privacy, copyright, and regulatory compliance in generative systems.

4.1 Mathematical Formalization

Model→Data attacks exploit information memorized within trained models to recover or expose private training data. Given a model $f_{\boldsymbol{\theta}}$ trained on $\mathcal{D}_{\text{train}}$, the objective G of the attacker is to extract training data set information $\boldsymbol{x}_{\text{info}}$ from the outputs or representations of the model $f_{\boldsymbol{\theta}}$:

$$\boldsymbol{x}_{\text{info}} = G(f_{\boldsymbol{\theta}}), \quad \text{s.t.} \quad \boldsymbol{x}_{\text{info}} \sim \mathcal{T}(\mathcal{D}_{\text{train}})$$

Here $\mathcal{T}(\mathcal{D}_{\text{train}})$ denotes the information about the training dataset that the attacker seeks to recover or infer from the model (e.g., specific samples, attributes, or membership signals). Constraints capture model access level and query limits. Different instantiations of G yield three major M$\mapsto$D families: *model inversion attack*, *membership inference attack*, and *training data extraction attack*.

4.1.1 Special Case I

Model Inversion Attacks Yang et al. (2025a), under such attacks, $\boldsymbol{x}_{\text{info}}$ is the reconstructed data $\hat{\boldsymbol{x}}$ that aligns with the distribution of $\mathcal{D}_{\text{train}}$. Given a target output y^*, the attacker defines an inversion objective function G as:

$$\begin{aligned}\hat{\boldsymbol{x}} = G(f_{\boldsymbol{\theta}}) &:= \arg\min_{\boldsymbol{x}} \big[\mathcal{L}_{\text{inv}}(f_{\boldsymbol{\theta}}(\boldsymbol{x}), y^*) + \lambda\,\mathcal{R}(\boldsymbol{x})\big], \\ \text{s.t.} \quad & \hat{\boldsymbol{x}} \sim \mathcal{T}(\mathcal{D}_{\text{train}}).\end{aligned}$$

where := denotes definition, $\mathcal{L}_{\text{inv}}$ enforces output consistency with the target y^*, and $\mathcal{R}$ regularizes the realism of reconstructed samples. Both white-box (gradient-based) and black-box (API-query) settings can reveal sensitive attributes or even approximate original training samples. Here, y^* denotes the target output (label or response) with respect to which the attacker seeks an input $\hat{\boldsymbol{x}}$ whose model prediction matches y^*, and $\mathcal{T}(\mathcal{D}_{\text{train}})$ represents the distribution of the training dataset.

4.1.2 Special Case II

Membership Inference Attacks (MIA) Hu et al. (2022a) determine a a binary membership signal $\boldsymbol{x}_{\text{info}}$ whether a specific data sample $\boldsymbol{x}^*$ was part of the training dataset $\mathcal{D}_{\text{train}}$. Formally, the attacker designs a discriminator g (or classifier) that takes the model $f_{\boldsymbol{\theta}}$ outputs or representations to predict whether $\boldsymbol{x}^*$ belongs to $\mathcal{D}_{\text{train}}$. The objective of G can be written as:

$$\begin{aligned}g^* = G(f_{\boldsymbol{\theta}}) &:= \arg\min_{g}\ \mathcal{L}_{\text{mem}}\big(g(f_{\boldsymbol{\theta}}(\boldsymbol{x}^*)),\, m^*\big), \\ \text{s.t.} \quad & g(f_{\boldsymbol{\theta}}(\boldsymbol{x}^*)) \in [0,1].\end{aligned}$$

where $\mathcal{L}_{\text{mem}}$ denotes the membership classification loss (e.g., binary cross-entropy), and $m^* \in \{0,1\}$ is the ground-truth membership label indicating whether the target sample $\boldsymbol{x}^*$ belongs to the training dataset. The trained discriminator outputs $g(f_{\boldsymbol{\theta}}(\boldsymbol{x}^*))$ lies in $[0,1]$ and is defined with respect to the training-data information $\mathcal{T}(\mathcal{D}_{\text{train}})$.

4.1.3 Special Case III

Training Data Extraction Attacks Xu et al. (2024a) aim to induce a generative model to directly reproduce private samples or fragments from its training data. Given a generative model $f_{\boldsymbol{\theta}}$ that defines a conditional distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{q})$ over outputs given a query or prompt $\boldsymbol{q}$, the attacker constructs an extraction operator E that interacts with $f_{\boldsymbol{\theta}}$ to recover samples consistent with the training-data information $\mathcal{T}(\mathcal{D}_{\text{train}})$.

In this case, $\boldsymbol{x}_{\text{info}}$ corresponds to the generated samples that the attacker extracts from the model, in this case, denotes as $\tilde{\boldsymbol{x}}$, which are expected to align with the training data distribution $\mathcal{T}(\mathcal{D}_{\text{train}})$. The objective of G can be expressed as:

$$\begin{aligned}E^* = G(f_{\boldsymbol{\theta}}) &:= \arg\min_{E}\ \mathbb{E}_{\tilde{\boldsymbol{x}} \sim f_{\boldsymbol{\theta}}(\cdot|E)}\big[\mathcal{L}_{\text{rec}}\big(\tilde{\boldsymbol{x}},\, \mathcal{T}(\mathcal{D}_{\text{train}})\big)\big], \\ \text{s.t.} \quad & \tilde{\boldsymbol{x}} = f_{\boldsymbol{\theta}}(\cdot \mid E^*) \approx \mathcal{T}(\mathcal{D}_{\text{train}}).\end{aligned}$$

Here, $\approx$ denotes approximately equal to, E denotes the attacker's extraction operator (e.g., a query generator, decoding policy, or sampling strategy) that issues prompts or queries to $f_{\boldsymbol{\theta}}$. Given a model $f_{\boldsymbol{\theta}}$, the attacker can obtain a $\tilde{\boldsymbol{x}}$ through E, which is approximately equal to certain information contained in $\mathcal{T}(\mathcal{D}_{\text{train}})$; $\mathcal{L}_{\text{rec}}$ measures reconstruction fidelity or semantic similarity between generated outputs $\tilde{\boldsymbol{x}}$ and the target training information $\mathcal{T}(\mathcal{D}_{\text{train}})$. In practice, E may operate *targetedly*—optimizing queries toward a specific sample $\boldsymbol{x}^*$ or identifying memorized content associated with known attributes—or *untargetedly*, by probing the model to elicit any memorized fragments through repeated sampling.

Access assumptions The signals available to the adversary vary with the threat model: *(a) White-box access Fredrikson et al. (2015):* full gradients or hidden activations can be exploited to optimize reconstructions directly. *(b) Black-box access:* the adversary can only interact with the model through its outputs. Two common variants are: (b.1) probability access Zhang et al. (2020), where softmax scores or logits are returned and provide richer information for inversion; and (b.2) label-only Kahla et al. (2022), where only the top-1 predicted class is visible, making query efficiency critical.

4.2 Taxonomy and Techniques of Model→Data

4.2.1 Model Inversion Attacks

Model inversion attacks (MIAs) aim to reconstruct sensitive training data—or its representative features—from a model's accessible information, such as outputs, gradients, or embeddings. Recent studies Dibbo (2023); Yang et al. (2025a) have established MIAs as one of the core privacy threats linking models back to their data, with diverse forms depending on the attacker's objective, access, and prior knowledge. Below we outline these key dimensions and the main technical paradigms observed across modalities. *(a) Target Type.* Depending on the reconstruction granularity, attacks may operate at the *instance level*, recovering individual samples Fredrikson et al. (2015; 2014); the *class level*, reconstructing category prototypes or feature representations Hitaj et al. (2017); or the *distribution level*, approximating the overall data manifold through semantic or statistical priors Chen et al. (2021b). *(b) Access Level.* Depending on access, white-box settings reveal gradients or hidden activations Zhang et al. (2020); Hu et al. (2023a); Wei et al. (2024b;a), black-box settings expose only logits or labels Kahla et al. (2022); Hu et al. (2023b), and mixed cases exploit side channels or shared embeddings Chanpuriya et al. (2021). *(c) Prior Knowledge.* Attackers may leverage auxiliary in-domain information memorized by the model itself Carlini et al. (2019), or employ synthetic priors—such as noise-based reconstructions or randomly generated queries—to approximate the target label data distribution Truong et al. (2021); Tramèr et al. (2016).

Across various settings, methods can be broadly categorized into five families. (a) *Optimization-based inversion* reconstructs inputs by maximizing confidence or minimizing feature discrepancy on target labels, often regularized by perceptual or total-variation priors Zhang et al. (2020); Mahendran & Vedaldi (2015). (b) *Learning-based inversion* further trains up-convolutional networks to directly map feature representations back to images Dosovitskiy & Brox (2016). (c) *Generative inversion* Wang et al. (2021a) employs conditional GANs or variational inference to sample plausible reconstructions. (d) *Representation-space inversion* Tragoudaras et al. (2025) decodes intermediate embeddings into the input domain, while (e) *prompt- or explanation-guided inversion* Morris et al. (2023); Zhao et al. (2021) leverages logits, gradients, or attribution maps to refine reconstruction quality.

While early work focused on vision, similar mechanisms now extend to text Morris et al. (2023), graphs Zhou et al. (2023), time series, and medical signals Subbanna et al. (2021); Ghimire et al. (2018). Diffusion models exhibit both data extraction Carlini et al. (2023) and prompt-level inversion behaviors Mahajan et al. (2024). At the foundation scale, increasing capacity amplifies memorization Carlini et al. (2021), enabling instance-level leakage even through limited-access interfaces Dibbo (2023); Carlini et al. (2022). Model inversion can target single samples, classes, or distributions, exploiting gradients, logits, or representations to reverse the data–model mapping Dibbo (2023); Wei et al. (2025).

4.2.2 Membership Inference Attacks

Membership inference attacks aim to determine whether a particular sample—or a collection of samples—was included in a model's training data set. By exploiting prediction behavior, hidden activations, or gradients, these attacks expose whether the model *memorize* specific data, revealing data participation information and thus breaching data privacy. Current approaches can be broadly grouped into four families: (a) output-based black-box attacks, (b) internal-signal and white-box attacks, (c) data-extraction–driven leakage, and (d) domain- or system-specific extensions. *(a) output-based black-box attacks.* These methods rely only on model outputs such as likelihoods, confidence scores, or generated text. Prompt- and perturbation-based attacks Fu et al. (2025) probe stability in next-token probabilities or generations to separate members from non-members. Likelihood-based detection at the *dataset/corpus* level is re-evaluated and formalized by *Min-K% Inference* Maini et al. (2024). In retrieval-augmented generation (RAG) systems, membership can be inferred from the semantic similarity and perplexity between retrieved knowledge and generated text, revealing private entries in the external database Li et al. (2025b). Complementarily, black-box provenance detection frameworks for LLMs—e.g., *DPDLLM*—identify whether text likely appeared in pre-training without logit access Zhou et al. (2024). *(b) Internal-Signal and White-Box Attacks.* When gradients, weights, or activations are observable, stronger inference becomes possible. Gradient- and parameter-based methods Pang et al. (2023); Suri et al. (2024) leverage differential signals to expose training membership, while neuron-level attribution analysis (*Unveiling the Unseen*) Li et al. (2024a) identifies internal activations correlated with membership cues. Statistical testing frameworks such as *Low-Cost High-Power Membership Inference Attacks* (RMIA) Zarifzadeh et al. (2023) further improve sensitivity and robustness under limited reference models. For large language models, *memTrace* Makhija et al. (2025) extracts membership signals from hidden-state dynamics and attention patterns. At the other end of the access spectrum, *OSLO* Peng et al. (2024) demonstrates that even label-only interfaces enable high-precision inference via transfer-based adversarial perturbations. Finally, explainability mechanisms themselves can act as side channels, as attribution maps and confidence changes from explanators reveal membership information Liu et al. (2024b). *(c) Data extraction and inversion leakage.* These attacks reveal a continuum between membership inference and explicit data reconstruction. Diffusion and language models can directly regenerate training samples through repeated prompting or sampling Carlini et al. (2023; 2021), showing that memorization and membership inference are deeply entangled. *(d) Domain-specific and system-level extensions.* Beyond standard classifiers, membership inference has been explored in graph contrastive learning, recommender systems, and biometric recognition Wang & Wang (2024); Zhong et al. (2024). Comprehensive evaluations DeAlcala et al. (2024); Zhu et al. (2024a) identify key influencing factors across centralized and federated settings, while information-theoretic and learning-based calibration analyses Zhu et al. (2025b); Shi et al. (2024) provide finer-grained quantification of leakage. Membership inference spans a continuum from black-box querying to gradient-level forensics, connecting with data extraction, inversion, and unlearning analysis. Across modalities and model scales, these studies collectively show that memorization remains a fundamental — and quantifiable — privacy risk in modern machine learning.

4.2.3 Training Data Extraction Attacks

Training data extraction aims to recover verbatim training data from deployed models and can be grouped by the adversary's access level and manipulation capability. Existing work mainly focuses on *query-based extraction*, which elicits memorized content via prompt engineering and sampling artifacts. Black-box adversaries craft prompts or exploit sampling errors to trigger memorized responses. Sequence-level studies Xu et al. (2024a) show that shorter prefixes and larger models tend to leak more. Recent work by Nasr et al. Nasr et al. (2025) introduces two scalable attacks—*divergence* and *finetuning*——that enable large-scale recovery of proprietary training data even under restricted, publicly accessible interfaces. Building on this direction, More et al. More et al. (2024) examine more realistic adversarial settings, showing that prompt sensitivity, access to multiple checkpoints, and downstream tasks can amplify extraction risks, revealing a stronger composite adversary that better captures real-world threat conditions. Parallel efforts extend these attacks to generative diffusion models: Carlini et al. Carlini et al. (2023) demonstrate training-image extraction from diffusion models, and Chen et al. Chen et al. (2024) further propose *SIDE*, a surrogate-conditioning framework enabling recovery even from unconditional models.

4.3 Defensive Techniques.

4.3.1 Defenses against Model Inversion Attacks

These defenses aim to protect the privacy and confidentiality of the data used to train or query a model. They can be broadly categorized into three complementary directions: *(a) Training-Time Privacy Regularization.* The dominant approach is differentially private stochastic gradient descent (DP-SGD) Abadi et al. (2016), which injects calibrated noise into gradient updates to bound each sample's contribution. While DP provides formal privacy guarantees, it often degrades accuracy. Complementary strategies introduce implicit regularization: information-bottleneck–inspired methods such as bilateral dependency optimization Peng et al. (2022) constrain representations to retain only task-relevant features, and stochastic mechanisms such as dropout Srivastava et al. (2014) mitigate overfitting and implicitly reduce memorization. *(b) Inference-Time Output Obfuscation.* Since most inversion attacks rely on observable outputs, these defenses modify predictions to conceal exploitable signals. Label smoothing Müller et al. (2019) reduces confidence gaps across classes and improves calibration, while adversarial regularization Wen et al. (2021) or randomized post-processing of logits weakens gradient-based reconstruction cues. *(c) Post-Deployment Detection and Perturbation Frameworks.* Runtime defenses focus on detecting inversion-like queries or embedding irreversible transformations into model internals. Semantic perturbation–based frameworks Zhu et al. (2024b) analyze query embeddings and behavioral consistency, introducing statistical signatures that help distinguish or distort inversion-derived surrogates.

4.3.2 Defenses against Membership Inference Attacks

These defenses aim to eliminate the statistical gap between members and non-members observable from model outputs, gradients, or representations. Existing studies fall into four major categories: *(a) unlearning and token-level mitigation*, *(b) ensemble and distillation strategies*, *(c) noise injection and regularization*, and *(d) generative and adversarial training.*

(a) Unlearning and Token-Level Mitigation. Selective unlearning treats memorized content differently from general knowledge. *Tokens for Learning, Tokens for Unlearning* Tran et al. (2025) jointly optimizes learning and unlearning objectives by categorizing tokens into "hard" and "memorized," reducing membership leakage with minimal impact on language modeling performance. Other unlearning methods apply targeted forgetting or data editing to erase memorized content while preserving task utility. *(b) Ensemble and Distillation Strategies.* These methods aggregate or transfer knowledge across multiple models to dilute membership signals. Multi-teacher and repeated distillation frameworks Zheng et al. (2021); Shejwalkar & Houmansadr (2021) transfer softened or masked predictions from teacher models to students, mitigating overconfident behaviors and enabling tunable privacy–utility trade-offs. Ensemble-based defenses such as MIAShield Jarin & Eshete (2023) mitigate membership inference by preemptively excluding models trained on the queried sample, thereby eliminating strong membership signals while preserving utility. *(c) Noise Injection and Regularization.* A widely used direction is to perturb training or inference signals to blur member/non-member distinctions. Noise injection methods (e.g., Weighted Smoothing Tan et al. (2023)) adaptively add perturbations to high-risk samples, while regularization-based defenses (e.g., MIST Li et al. (2024b) and NeuGuard Xu et al. (2022)) constrain model representations or neuron activations to reduce membership inference vulnerability. Graph perturbation Wang et al. (2023a) obscures membership signals by injecting noise into graph structures, while enhanced Mixup Chen et al. (2021c) and weight pruning Wang et al. (2021b) regularize models to reduce overfitting and memorization. *(d) Generative and Adversarial Training.* Generative defenses leverage GAN- or VAE-based frameworks to generate synthetic data to train or regularize generative models, in order to obscure membership signals while preserving utility Hu et al. (2022b); Mukherjee et al. (2021); Yang et al. (2023). Digestive neural networks Lee et al. (2021) sanitize shared gradients in federated settings, and adversarial regularization Nasr et al. (2018) jointly trains a classifier and adversary to produce membership-resistant representations.

4.3.3 Defenses against Training Data Extraction Attacks

Large language and diffusion models may memorize and expose sensitive training examples through overfitting or sampling. Defenses, therefore, aim to suppress memorization, distort membership signals, or detect exposed content. Existing works can be grouped into four complementary directions. *(a) Training-Time Regularization and Noise.* These defenses modify the optimization process to ensure similar behavior between members and non-members. Differentially private fine-tuning (e.g., DP-SGD or DP-Adam) Du et al. (2025) provides formal privacy guarantees against data leakage and, when combined with low-rank adaptation (LoRA), achieves a favorable privacy–utility trade-off. *(b) Architectural and Ensemble Isolation.* Re-architecting models to decouple knowledge across data subsets prevents any single component from over-memorizing, as demonstrated by SELENA Tang et al. (2022), which trains multiple sub-models on overlapping subsets and uses self-distillation to align their behaviors. *(c) Query- and Output-Level Perturbation.* Post-training defenses perturb model queries or responses to obfuscate membership signals. QUEEN Chen et al. (2025a) adaptively perturbs sensitive queries and reverses gradients to corrupt extraction attempts in model-stealing scenarios. Beyond query perturbation, output watermarking offers another form of post-generation modification. Panaitescu et al. Panaitescu-Liess et al. (2025) show that output watermarking can significantly reduce the probability of verbatim memorization, thereby preventing copyrighted text generation. *(d) Memorization Detection and Auditing.* Rather than suppressing leakage, these defenses identify and monitor it. Diffusion-model audits Wen et al. (2024) detect memorized samples via prompt-conditioned prediction analysis, while LLM auditing frameworks such as ContextLeak Choi et al. insert canaries or triggers to trace data exposure during fine-tuning and in-context learning.

5 Model→Model (M→M)

5.1 Model Extraction Attacks

Model Extraction Attacks (MEAs) Liang et al. (2024) pose a critical threat to the confidentiality and intellectual property (IP) of machine learning models, particularly in the context of Machine Learning as a Service (MLaaS)Kesarwani et al. (2018). In MEAs, an attacker with black-box access queries a proprietary model and uses the responses to train a substitute that replicates the victim model's functionality, architecture or parameters, etc. This surrogate can approximate not only the model's decision boundaries, but sometimes its architecture or parameters—undermining commercial value and IP protection.

MEAs can be broadly classified into two main categories: *(1) Functionality Stealing Truong et al. (2021)*: the adversary aims to replicate the prediction performance of the victim model, producing a substitute that yields consistent outputs. Depending on the availability of data, functionality stealing can be further divided into: *(a) Data-based Model Extraction (DBME)*, where attackers leverage knowledge of the training dataset used for training the target victim model or a surrogate dataset to query the victim model and distill its knowledge; and *(b) Data-free Model Extraction (DFME)*, where no prior knowledge of the target victim model's training data is known and the synthesis of the attacker's query data is iteratively refined using the victim model's outputs as feedback. *(2) Architecture Stealing Rolnick & Kording (2020)*: the goal is to infer the internal design of the victim model, such as its layer structure or hyperparameters. Unlike functionality stealing, which focuses on prediction performance, architecture stealing targets the proprietary network design itself, enabling adversaries to reconstruct or optimize their own models.

5.2 Mathematical Formalization

Attacker's Knowledge and Objective. In a model extraction attack, the adversary interacts with a victim model $V(\boldsymbol{x};\boldsymbol{\theta}_V)$ solely through its API. By submitting a set of queries $X=\{\boldsymbol{x}_i\}_{i=0}^{i=m}$, attacker receives corresponding outputs:

$$y_i = V(\boldsymbol{x}_i;\boldsymbol{\theta}_V), \quad i=1,\dots,m,$$

which may be either probability vectors (*soft-labels*) or top-1 predictions (*hard-labels*), depending on the API configuration. Using the collected pairs $\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{i=m}$, the attacker then trains a substitute (clone) model $C(\boldsymbol{x};\boldsymbol{\theta}_C)$ with the goal of reproducing the predictive behavior of V. Formally, the surrogate model's parameters are learned by solving:

$$\boldsymbol{\theta}_C^{\star} \in \arg\min_{\boldsymbol{\theta}_C} \sum_{i=1}^{m} \mathcal{L}(\boldsymbol{x}_i, y_i, \boldsymbol{\theta}_C)$$

where $\mathcal{L}$ is an appropriate loss function (e.g., distillation loss or cross-entropy loss). Based on whether prior information of training data is available, MEAs can be categorized into *data-based* (DBME) and *data-free* (DFME) approaches.

Defender's Knowledge and Objective Defender's aim is to maintain the victim model's accuracy on its in-distribution (ID) dataset while simultaneously degrading the utility of any cloned model trained through the extraction attack. In practice, the defender operates under limited knowledge: the precise attack strategy, the architecture of the clone model, and whether a query is benign or adversarial are typically unknown. A common assumption is that adversarial queries originate from out-of-distribution (OOD) data Kariyappa & Qureshi (2020), since the original training set is private and rarely exposed to API users. Nevertheless, effective defenses should also remain applicable when attackers have access to in-distribution queries, ensuring robustness across both DBME- and DFME-style attacks.

5.3 Taxonomy and Techniques of MEAs

We categorize model extraction attacks into two main types — *data-based* and *data-free* — which differ in whether the attacker has prior access to the victim model's training data.

5.3.1 Data-Based Attacks

Data-based extraction begins from public or domain-related inputs, leveraging semantic priors for faster convergence and higher sample utility.

(a) One stream of work focuses on query selection and augmentation. The seminal JBDA approach Papernot et al. (2017) introduced Jacobian-based dataset augmentation, training surrogate model from black-box label outputs to approximate decision boundaries. *Copycat CNN* Correia-Silva et al. (2018) shows that using non-problem-domain natural images can replicate the functionality of target models. *ActiveThief* Pal et al. (2020) integrates K-center and active learning to select uncertain samples, achieving higher agreement with fewer queries. Augmentation and ensembles can further increase the informativeness of each query. For example, *Army of Thieves* Jindal et al. (2024) employs ensemble-based consensus entropy and label disagreement to guide query selection. Meanwhile, *AugSteal* Gao et al. (2024b) combines Grad-CAM-based data filtering and MPCL active selection with a FusionAug module (GridMix + MulAug) to enhance functional similarity under hard-label constraints. Finally, *MARICH* Karmakar & Basu (2024) matches the victim model output distribution via entropy- and divergence-based objectives, achieving high fidelity to the victim model's performance.

(b) A second stream explicitly probes the decision boundary of the victim model. *CloudLeak* Yu et al. (2020) employs adversarial and active-learning–based queries to explore regions near classification boundaries, while *BEST* Li et al. (2022) refines this idea with entropy-driven uncertain examples to capture both accuracy and robustness regions. *SPSG* Zhao et al. (2024b) leverages superpixel segmentation and low-variance gradient estimation to approximate boundary information efficiently under limited queries. Other approaches, such as *InverseNet* Gong et al. (2021) and *LOKT* Nguyen et al. (2024a), reconstruct the victim's data distribution respectively through input inversion and label-space generative transfer.

(c) A third category of attacks leverages extra signals (e.g., explanations) to enhance the effectiveness of model extraction. Explanation-based attacks such as *XaMEA* Yan et al. (2023) exploit explanation signals (e.g., saliency maps) to enhance surrogate fidelity, while *PtbStolen* Zhang et al. (2023a) steals encoder representations via feature-vector matching on perturbed samples. At the system level, even when only hard labels are accessible, query-based knockoff and Jacobian-augmentation attacks remain feasible Tramèr et al. (2016). Similarly, limited access to RNN hidden states can suffice for replication Takemura et al. (2020).

Foundation models (Data-Based) For LLMs, *in-context* imitation attacks Li et al. (2024c) demonstrate that even without gradient access, medium-sized models can reproduce specialized abilities such as code summarization and synthesis by querying black-box APIs with carefully designed prompts. Here, traditional datasets are replaced by curated prompt sets and task suites, indicating that prompt programming itself can function as a data-based extraction strategy. Together with earlier observations on boundary probing and attribution-guided attacks, these results highlight that restricting access to the API interface alone is insufficient to prevent high-fidelity cloning.

5.3.2 Data-Free Attacks

Data-free model extraction (DFME) synthesizes Q without access to in-distribution data. Then, the generation strategies accelerated progress: *MAZE* Kariyappa et al. (2021a) guides a generator toward regions of maximal model disagreement via zeroth-order gradient estimation; and *DFMS-HL* Sanyal et al. (2022) adapts to hard-label constraints with class-diversity regularization and adversarial alignment terms. *DFMS-DG* He et al. (2024) utilizes denoising diffusion GANs to generate diverse, high-quality samples that improve clone model accuracy even against adversarially trained targets. DFCL-APIs Yang et al. (2024a) extends the study of model extraction to the continual learning paradigm.

While generator-driven methods remain central to DFME, recent approaches enhance them with explicit optimization and sample selection strategies. For example, *ES Attack* Yuan et al. (2022) and *Truong et al.* Truong et al. (2021) iteratively refine synthetic queries through alternating estimation and synthesis steps to distill the victim model, while *DFHL-RS* Yuan et al. (2024) generates high-entropy examples near decision boundaries and reuses them to reduce costs under hard-label constraints. Because DFME is highly budget-sensitive, several methods emphasize query efficiency: *IDEAL* Zhang et al. (2022) decouples generation from distillation, requiring only one query per synthetic sample, whereas E^3 Zhu et al. (2025a) improves efficiency via language-guided sample selection, multi-resolution training, and temperature tuning, achieving comparable fidelity with merely about 0.5% of the queries needed by conventional GAN-based methods; and *DualCOS* Yang et al. (2024b) incorporates active sampling and disagreement-based objectives to maximize sample reuse. Under label-only APIs, *QUDA* Lin et al. (2023) employs deep reinforcement learning with weak generative priors; and *DisGUIDE* Rosenthal et al. (2023) drives divergence in clone outputs to increase informativeness while reducing query counts. DFME also extends beyond classification: *Yue et al.* Yue et al. (2021) demonstrate that generating synthetic queries can effectively replicate victim recommender models.

Foundation models (Data-Free) DFME is no longer vision-only. In language, *Lion* Jiang et al. (2023b) employs adversarial knowledge distillation with a feedback loop that generates hard instructions to efficiently distill GPT-style models using only 70K queries, achieving ChatGPT-level performance with minimal supervision. Random or task-agnostic queries can also suffice for BERT-based APIs Krishna et al. (2020). Beyond text, *LCA* Shao et al. (2025a) leverages Stable Diffusion's latent prior for adversarial query synthesis, combining latent augmentation with membership-aware sampling to produce high-utility prompts in text and multimodal settings.

In both data-based and data-free settings, recent studies Jagielski et al. (2020); He et al. (2021); Dai et al. (2023) have proposed extraction methods that generalize well across domains and remain effective even with a limited number of queries. Foundation models further amplify these risks because their prompt interaction interfaces and few-shot generalization make high-fidelity cloning more practical. These developments highlight the need for defense-in-depth strategies—including dynamic output perturbation, query-pattern monitoring, attribution filtering, and robust watermarking—that are adapted to the attacker's query regime.

5.4 Defensive Techniques.

Existing model extraction defenses can be broadly categorized into two classes: *prevention defenses*, which actively degrade the extraction process, and *verification/detection defenses*, which passively monitor or verify whether model extraction has occurred.

5.4.1 Model extraction prevention defenses (active defenses)

These defenses aim to reduce the value of queries or limit the fidelity of stolen models. Representative strategies include:

(a) Model output perturbation. A primary line of work modifies API outputs to reduce the information available to an attacker. Techniques include selective output perturbation, response filtering, or adaptive shaping

that perturbs outputs only for abnormal queries while preserving accuracy on legitimate ones. Examples include *AdvFT* Zhang et al. (2024), which perturbs feature representations of out-of-distribution (OOD) queries; *CIP* Zhang et al. (2023b), which combines energy-based OOD scoring with selective poisoning and traceable watermarking; *AMAO* Jiang et al. (2023a), which integrates adversarial training, adaptive outputs, and embedded watermarks; *ModelGuard* Tang et al. (2024), which adaptively optimizes perturbations to balance information leakage and prediction utility; and *Noise Transition Matrices* Wu et al. (2024), which inject lightweight structured noise.

(b) Training-time defenses. Some defenses alter the model itself during training to make model extraction harder. Defensive training strategies Wang et al. (2023b); Hong et al. (2024) inject robustness by learning causal or distributionally robust representations that hinder surrogate learning. Architectural modifications include *DNF* Luan et al. with early exits, and *InI* Guo et al. (2023) which dynamically isolates suspicious queries to block gradient-based extraction. *MeCo* Wang et al. (2023b) further employs distributionally robust defensive training with a data-dependent perturbation generator to resist data-free extraction. AAUG Wang et al. (2024a) which introduces an attack-aware and uncertainty-guided training objective to reduce the accuracy of stolen models, and Beowulf Gong et al. (2024) introduces adversarial dummy classes during training, which reshape decision boundaries to mislead surrogate models and degrade the fidelity of surrogate models. RL-based meta-policies Orekondy et al. (2019) adapt label shaping dynamically, while ensemble defenses Kariyappa et al. (2021b) diversify decision boundaries to resist approximation. Active watermarking Wang et al. (2024b) fine-tunes models to embed probabilistic signals that actively degrade the performance of cloned models.

5.4.2 Model extraction verification/detection defenses (passive defenses)

These defenses do not prevent model extraction directly, but provide evidence or detection signals.

(a) Query-level monitoring and anomaly detection. Detection-based methods identify extraction attempts in real time. Statistical defenses such as *PRADA* Juuti et al. (2019) analyze query distributions, while *SEAT* Zhang et al. (2021) and *HODA* Sadeghzadeh et al. (2023) leverage query similarity to detect structured exploration. More advanced anomaly detectors include *SAME* Xie et al. (2024), which uses autoencoder reconstruction error, and adaptive detection-and-response methods Kariyappa & Qureshi (2020). Although effective, these methods can be bypassed if adversaries disguise their query distribution Azizmalayeri et al. (2022).

(b) Verification-based watermarking. Another class of defenses embeds verifiable signals into models so that ownership can be established after suspected theft. Earlier trigger-set or boundary-based approaches Adi et al. (2018); Zhang et al. (2018) and subsequent schemes such as *DeepSigns* Darvish Rouhani et al. (2019) enable IP owners to query a suspect model and check for expected responses. These methods include both training-time and API-level watermarking, providing strong ownership verification but offering no direct prevention against model extraction. For generative models, *WDM* Peng et al. (2023) embeds watermarks into diffusion U-Nets, while adversarial or lexical watermarking He et al. (2022) inserts detectable linguistic or statistical patterns that later facilitate ownership verification. Representative approaches such as *EWE* Jia et al. (2021) and *DAWN* Szyller et al. (2021) further embed persistent signals that survive in surrogate models, thereby enabling reliable ownership tracing. More recently, *LIDet* Zhao et al. extends this line of work to large language models, using text-level watermarks to detect intellectual property infringement in suspect LLMs under black-box access.

In summary, model extraction defenses fall into two complementary categories. *Prevention defenses* actively interfere with the attacker by shaping API outputs, or modifying training objectives, thereby lowering the fidelity or usability of stolen models. *Verification and detection defenses*, in contrast, do not stop extraction but provide monitoring signals (e.g., anomaly detection on queries) or verifiable marks for post-hoc attribution. In practice, robust protection often combines both: prevention raises the attack cost and reduces the value of extraction, while verification provides the evidence and legal traceability once theft has occurred.

5.5 From Functional Imitation to Structural Theft

So far, Model→Model ($M \rightarrow M$) discussion has focused on model *functional* cloning, where the goal is to reproduce the victim's input–output behavior. More recent attacks target *structural* recovery, inferring model architectures, parameters, or training settings. This shift changes the objective (from output agreement to structural fidelity) and increases risk: recovering model internals lowers the cost of subsequent attacks and can defeat defenses that rely solely on restricting API outputs.

Structural and Parametric Extraction *(a) From structural inference to architecture reconstruction.* Early studies demonstrated that black-box outputs can leak model architecture and training attributes. Rolnick et al. Rolnick & Kording (2020) exploited the piecewise linearity of ReLU networks to analytically reconstruct their layer topology and parameters, yielding functionally equivalent models. *(b) Practical weight recovery via learning and analysis.* Jagielski et al. Jagielski et al. (2020) combined surrogate imitation with direct analytical recovery, achieving near-exact weight reconstruction for multi-layer ReLU networks using only a few thousand queries. Carlini et al. Carlini et al. (2020) reformulated parameter recovery as a differential-cryptanalysis problem, highlighting the feasibility of fine-grained extraction under black-box access. *(c) Partial reconstruction in large-scale production models.* Extending to commercial settings, Carlini et al. Carlini et al. (2024b) showed that even limited API feedback (e.g., log probabilities or logit bias) suffices to reconstruct key components of large-scale LLMs with high fidelity and modest cost, underscoring practical risks for deployed models.

6 Discussion and Future Directions

The foundation model era has fundamentally reshaped the landscape of AI security. As models scale in capability, modality, and accessibility, the boundaries between data and models have blurred—making it increasingly difficult to protect one without considering the other. Our unified closed-loop taxonomy highlights that security threats are not isolated phenomena but are interconnected spanning data-to-data, data-to-model, model-to-data, and model-to-model interactions. Moving forward, safeguarding AI systems requires integrated, adaptive, and theoretically principled approaches that can evolve alongside the rapid advancement of foundation models.

6.1 Toward Adaptive and Continuous Defenses

A major challenge across all four interaction pathways is the static nature of current defenses. Most countermeasures are designed with fixed assumptions about attacker behavior, dataset composition, or model deployment environments. However, the rise of generative and adaptive adversaries—who exploit model updates, fine-tuning interfaces, or evolving prompt strategies—demands *continuous security mechanisms.* Future research should focus on developing self-adaptive defenses capable of monitoring model–data interactions in real time, detecting distributional shifts, and autonomously mitigating emerging threats. This may involve online learning for threat detection, continual robustness calibration, and the use of meta-learning or reinforcement learning to adapt defense strategies dynamically.

6.2 Unifying Theoretical and Practical Foundations

Another key direction lies in bridging the gap between theoretical security guarantees and real-world practicality. While many defenses rely on heuristics such as loss reweighting, output perturbation, or rule-based filtering, few provide provable robustness bounds or formal privacy guarantees. Establishing a theoretical foundation that unifies differential privacy, information theory, and robustness analysis will be essential for quantifying and constraining information leakage across data-model pathways. Additionally, future work should explore *formal trade-off analyses*—understanding how accuracy, alignment, and security interact, and how to optimize them jointly within limited computational and data budgets. This theoretical rigor will also facilitate regulatory and certification frameworks for trustworthy AI deployment.

6.3 Cross-Modal and Cross-System Security

The current security literature remains largely modality-specific, with methods developed for vision, text, or speech systems in isolation. Yet foundation models increasingly integrate multiple modalities, enabling new vulnerabilities that emerge from cross-modal transfer and alignment failures. Future studies should investigate security mechanisms that generalize across data types and architectures—building *cross-modal benchmarks* and *unified evaluation protocols* to measure consistency and robustness across diverse input and output spaces. Furthermore, foundation models are now components within larger ecosystems of agents, APIs, and retrieval systems, where model-to-model interactions create compound risks. Addressing this complexity will require system-level perspectives that model the cascading effects of local vulnerabilities throughout interconnected AI pipelines.

6.4 Balancing Security, Utility, and Alignment

Ensuring robust security inevitably introduces tensions with model utility and alignment. Stronger defenses—such as randomized responses, output filtering, or adversarial fine-tuning—can degrade model performance or restrict generalization. Conversely, improving alignment through fine-tuning may inadvertently reduce robustness to adversarial manipulation. Future research should thus aim to design *joint optimization objectives* that harmonize these competing factors, ensuring models remain both safe and functional under realistic deployment conditions. Exploring energy-efficient and scalable defense architectures will also be crucial to make these mechanisms viable for industry-scale systems.

6.5 Collaborative Ecosystem and Benchmark Development

Finally, progress in AI security will depend on a collaborative ecosystem that unites academic, industrial, and policy communities. The field urgently needs open, standardized benchmarks for evaluating attacks and defenses, spanning all data-model interaction types and multiple modalities. Beyond algorithmic innovation, robust system design should integrate secure hardware, privacy-preserving computation, and transparent governance frameworks to ensure accountability. Publicly available datasets, shared risk assessments, and red-teaming initiatives will further help in identifying systemic vulnerabilities and advancing collective resilience.

6.6 Summary

The path toward secure and trustworthy AI in the foundation model era demands a paradigm shift—from isolated, reactive defenses to holistic, adaptive, and theoretically grounded systems that secure both data and models in tandem. Our unified closed-loop taxonomy provides a conceptual foundation for this shift, revealing how vulnerabilities propagate across data–model interactions and where new defensive strategies must emerge. By fostering collaboration across technical, theoretical, and policy domains, the community can move toward a resilient AI ecosystem that is not only powerful but also secure, transparent, and aligned with human values.

7 Conclusion

In summary, this work establishes a unified perspective on machine learning security by framing the interplay between data and models through a closed-loop threat taxonomy. By organizing interactions along four fundamental directions—data-to-data, data-to-model, model-to-data, and model-to-model—our framework reveals how vulnerabilities and defenses are interconnected across the entire ML pipeline. This holistic view not only clarifies existing threat relationships but also provides a foundation for designing more generalizable and resilient defense strategies in the era of large-scale and foundation models.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Yossi Adi, Carsten Baum, Moustapha Cisse, Benny Pinkas, and Joseph Keshet. Turning your weakness into a strength: Watermarking deep neural networks by backdooring. In *27th USENIX security symposium (USENIX Security 18)*, pp. 1615–1631, 2018.

Daniel Alexander Alber, Zihao Yang, Anton Alyakin, Eunice Yang, Sumedha Rai, Aly A Valliani, Jeff Zhang, Gabriel R Rosenbaum, Ashley K Amend-Thomas, David B Kurland, et al. Medical large language models are vulnerable to data-poisoning attacks. *Nature Medicine*, 31(2):618–626, 2025.

Mohammad Azizmalayeri, Arshia Soltani Moakar, Arman Zarei, Reihaneh Zohrabi, Mohammad Taghi Manzuri, and Mohammad Hossein Rohban. Your out-of-distribution detection method is not robust! In *Advances in Neural Information Processing Systems*, 2022.

Mauro Barni, Franco Bartolini, Vito Cappellini, and Alessandro Piva. A dct-domain system for robust image watermarking. *Signal processing*, 66(3):357–372, 1998.

Mahbuba Begum and Mohammad Shorif Uddin. Digital image watermarking techniques: a review. *Information*, 11(2):110, 2020.

Fabrice Benhamouda, Akshay Degwekar, Yuval Ishai, and Tal Rabin. On the local leakage resilience of linear secret sharing schemes. In *Annual International Cryptology Conference*, pp. 531–561. Springer, 2018.

Eitan Borgnia, Valeriia Cherepanova, Liam Fowl, Amin Ghiasi, Jonas Geiping, Micah Goldblum, Tom Goldstein, and Arjun Gupta. Strong data augmentation sanitizes poisoning and backdoor attacks without an accuracy tradeoff. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3855–3859. IEEE, 2021.

Raphael Bost, Raluca Ada Popa, Stephen Tu, and Shafi Goldwasser. Machine learning classification over encrypted data. *Cryptology ePrint Archive*, 2014.

Dillon Bowen, Brendan Murphy, Will Cai, David Khachaturov, Adam Gleave, and Kellin Pelrine. Scaling trends for data poisoning in llms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 27206–27214, 2025.

Zikui Cai, Shayan Shabihi, Bang An, Zora Che, Brian R Bartoldson, Bhavya Kailkhura, Tom Goldstein, and Furong Huang. Aegisllm: Scaling agentic systems for self-reflective defense in llm security. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pp. 267–284, 2019.

Nicholas Carlini, Matthew Jagielski, and Ilya Mironov. Cryptanalytic extraction of neural network models. In *Annual international cryptology conference*, pp. 189–218. Springer, 2020.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2022.

Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning web-scale training datasets is practical. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 407–425. IEEE, 2024a.

Nicholas Carlini, Daniel Paleka, Krishnamurthy Dvijotham, Thomas Steinke, Jonathan Hayase, A Feder Cooper, Katherine Lee, Matthew Jagielski, Milad Nasr, Arthur Conmy, et al. Stealing part of a production language model. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 5680–5705, 2024b.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX security symposium (USENIX Security 23)*, pp. 5253–5270, 2023.

Tarek Chaalan, Shaoning Pang, Joarder Kamruzzaman, Iqbal Gondal, and Xuyun Zhang. The path to defence: A roadmap to characterising data poisoning attacks on victim models. *ACM Computing Surveys*, 56(7):1–39, 2024.

Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. A survey on adversarial attacks and defences. *CAAI Transactions on Intelligence Technology*, 6(1):25–45, 2021.

Sudhanshu Chanpuriya, Cameron Musco, Konstantinos Sotiropoulos, and Charalampos Tsourakakis. Deepwalking backwards: from embeddings back to graphs. In *International conference on machine learning*, pp. 1473–1483. PMLR, 2021.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramer, et al. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *Advances in Neural Information Processing Systems*, 37:55005–55029, 2024.

Canyu Chen, Baixiang Huang, Zekun Li, Zhaorun Chen, Shiyang Lai, Xiongxiao Xu, Jia-Chen Gu, Jindong Gu, Huaxiu Yao, Chaowei Xiao, et al. Can editing llms inject harm? In *Neurips Safe Generative AI Workshop 2024*.

Huajie Chen, Tianqing Zhu, Lefeng Zhang, Bo Liu, Derui Wang, Wanlei Zhou, and Minhui Xue. Queen: Query unlearning against model extraction. *IEEE Transactions on Information Forensics and Security*, 2025a.

Jian Chen, Xuxin Zhang, Rui Zhang, Chen Wang, and Ling Liu. De-pois: An attack-agnostic defense against data poisoning attacks. *IEEE Transactions on Information Forensics and Security*, 16:3412–3425, 2021a.

Pengpeng Chen, Yongqiang Yang, Dingqi Yang, Hailong Sun, Zhijun Chen, and Peng Lin. Black-box data poisoning attacks on crowdsourcing. In *IJCAI*, pp. 2975–2983, 2023.

Si Chen, Mostafa Kahla, Ruoxi Jia, and Guo-Jun Qi. Knowledge-enriched distributional model inversion attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16178–16187, 2021b.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

Yunhao Chen, Shujie Wang, Difan Zou, and Xingjun Ma. Extracting training data from unconditional diffusion models. *arXiv preprint arXiv:2410.02467*, 2024.

Zixuan Chen, Weikai Lu, Xin Lin, and Ziqian Zeng. Sdd: Self-degraded defense against malicious fine-tuning. *arXiv preprint arXiv:2507.21182*, 2025b.

Zongqi Chen, Hongwei Li, Meng Hao, and Guowen Xu. Enhanced mixup training: A defense method against membership inference attack. In *International Conference on Information Security Practice and Experience*, pp. 32–45. Springer, 2021c.

Jacob Choi, Shuying Cao, Xingjian Dong, and Sai Praneeth Karimireddy. Contextleak: Auditing leakage in private in-context learning methods. In *The Impact of Memorization on Trustworthy Foundation Models: ICML 2025 Workshop*.

Antonio Emanuele Cinà, Kathrin Grosse, Ambra Demontis, Sebastiano Vascon, Werner Zellinger, Bernhard A Moser, Alina Oprea, Battista Biggio, Marcello Pelillo, and Fabio Roli. Wild patterns reloaded: A survey of machine learning security against training data poisoning. *ACM Computing Surveys*, 55(13s): 1–39, 2023.

Jacson Rodrigues Correia-Silva, Rodrigo F Berriel, Claudine Badue, Alberto F De Souza, and Thiago Oliveira-Santos. Copycat cnn: Stealing knowledge by persuading confession with random non-labeled data. In *2018 International joint conference on neural networks (IJCNN)*, pp. 1–8. IEEE, 2018.

Ingemar J Cox, Joe Kilian, F Thomson Leighton, and Talal Shamoon. Secure spread spectrum watermarking for multimedia. *IEEE transactions on image processing*, 6(12):1673–1687, 1997.

Guohao Cui, Cihui Yang, and Jianyong Guo. Generative adversarial networks for rubber stamp extraction and removal. *Digital Signal Processing*, 146:104358, 2024.

Chengwei Dai, Minxuan Lv, Kun Li, and Wei Zhou. Meaeq: Mount model extraction attacks with efficient queries. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12671–12684, 2023.

Bita Darvish Rouhani, Huili Chen, and Farinaz Koushanfar. Deepsigns: An end-to-end watermarking framework for ownership protection of deep neural networks. In *Proceedings of the twenty-fourth international conference on architectural support for programming languages and operating systems*, pp. 485–497, 2019.

Daniel DeAlcala, Gonzalo Mancera, Aythami Morales, Julian Fierrez, Ruben Tolosana, and Javier Ortega-Garcia. A comprehensive analysis of factors impacting membership inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3585–3593, 2024.

Tali Dekel, Michael Rubinstein, Ce Liu, and William T Freeman. On the effectiveness of visible watermarks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2146–2154, 2017.

Sayanton V Dibbo. Sok: Model inversion attack landscape: Taxonomy, challenges, and future roadmap. In *2023 IEEE 36th Computer Security Foundations Symposium (CSF)*, pp. 439–456. IEEE, 2023.

Yi Ding, Guozheng Wu, Dajiang Chen, Ning Zhang, Linpeng Gong, Mingsheng Cao, and Zhiguang Qin. Deepedn: A deep-learning-based image encryption and decryption network for internet of medical things. *IEEE Internet of Things Journal*, 8(3):1504–1518, 2020.

Yinpeng Dong, Huanran Chen, Jiawei Chen, Zhengwei Fang, Xiao Yang, Yichi Zhang, Yu Tian, Hang Su, and Jun Zhu. How robust is google's bard to adversarial image attacks? *arXiv preprint arXiv:2309.11751*, 2023.

Alexey Dosovitskiy and Thomas Brox. Inverting visual representations with convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4829–4837, 2016.

Hao Du, Shang Liu, and Yang Cao. Can differentially private fine-tuning llms protect against privacy attacks? In *IFIP Annual Conference on Data and Applications Security and Privacy*, pp. 311–329. Springer, 2025.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1322–1333, 2015.

Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An {End-to-End} case study of personalized warfarin dosing. In *23rd USENIX security symposium (USENIX Security 14)*, pp. 17–32, 2014.

Jiri Fridrich. Symmetric ciphers based on two-dimensional chaotic maps. *International Journal of Bifurcation and chaos*, 8(06):1259–1284, 1998.

Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. Mia-tuner: adapting large language models as pre-training text detector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 27295–27303, 2025.

Dehong Gao, Yufei Ma, Sen Liu, Mengfei Song, Linbo Jin, Wen Jiang, Xin Wang, Wei Ning, Shanqing Yu, Qi Xuan, et al. Fashiongpt: Llm instruction fine-tuning with multiple lora-adapter fusion. *Knowledge-Based Systems*, 299:112043, 2024a.

Lijun Gao, Wenjun Liu, Kai Liu, and Jiehong Wu. Augsteal: Advancing model steal with data augmentation in active learning frameworks. *IEEE Transactions on Information Forensics and Security*, 2024b.

Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C Ranasinghe, and Surya Nepal. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th annual computer security applications conference*, pp. 113–125, 2019.

Jonas Geiping, Liam H Fowl, W Ronny Huang, Wojciech Czaja, Gavin Taylor, Michael Moeller, and Tom Goldstein. Witches' brew: Industrial scale data poisoning via gradient matching. In *ICLR*, 2021.

Simon Geisler, Tom Wollschläger, MHI Abdalla, Johannes Gasteiger, and Stephan Günnemann. Attacking large language models with projected gradient descent. In *ICML 2024 Next Generation of AI Safety Workshop*.

Sandesh Ghimire, Jwala Dhamala, Prashnna Kumar Gyawali, John L Sapp, Milan Horacek, and Linwei Wang. Generative modeling and inverse imaging of cardiac transmembrane potential. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 508–516. Springer, 2018.

Noah Golowich and Ankur Moitra. Edit distance robust watermarks via indexing pseudorandom codes. *Advances in Neural Information Processing Systems*, 37:20645–20693, 2024.

Xueluan Gong, Yanjiao Chen, Wenbin Yang, Guanghao Mei, and Qian Wang. Inversenet: Augmenting model extraction attacks with training data inversion. In *IJCAI*, pp. 2439–2447, 2021.

Xueluan Gong, Rubin Wei, Ziyao Wang, Yuchen Sun, Jiawen Peng, Yanjiao Chen, and Qian Wang. Beowulf: Mitigating model extraction attacks via reshaping decision regions. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 3838–3852, 2024.

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

Zhi-Hong Guan, Fangjun Huang, and Wenjie Guan. Chaos-based image encryption algorithm. *Physics letters A*, 346(1-3):153–157, 2005.

Jun Guo, Xingyu Zheng, Aishan Liu, Siyuan Liang, Yisong Xiao, Yichao Wu, and Xianglong Liu. Isolation and induction: Training robust deep neural networks against model stealing attacks. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 4178–4189, 2023.

Dongyoon Hahm, Taywon Min, Woogyeol Jin, and Kimin Lee. Unintended misalignment from agentic fine-tuning: Risks and mitigation. *arXiv preprint arXiv:2508.14031*, 2025.

Danny Halawi, Alexander Wei, Eric Wallace, Tony Tong Wang, Nika Haghtalab, and Jacob Steinhardt. Covert malicious finetuning: Challenges in safeguarding llm adaptation. In *Forty-first International Conference on Machine Learning*.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018.

Jianping He, Haichang Gao, and Yunyi Zhou. Enhancing data-free model stealing attack on robust models. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2024.

Luxi He, Mengzhou Xia, and Peter Henderson. What's in your" safe" data?: Identifying benign data that breaks safety. In *ICLR 2024 Workshop on Navigating and Addressing Data Problems for Foundation Models*.

Xuanli He, Lingjuan Lyu, Lichao Sun, and Qiongkai Xu. Model extraction and adversarial transferability, your bert is vulnerable! In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2006–2012, 2021.

Xuanli He, Qiongkai Xu, Lingjuan Lyu, Fangzhao Wu, and Chenguang Wang. Protecting intellectual property of language generation apis with lexical watermark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 10758–10766, 2022.

Briland Hitaj, Giuseppe Ateniese, and Fernando Perez-Cruz. Deep models under the gan: Information leakage from collaborative deep learning. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, pp. 603–618, 2017.

Ziming Hong, Zhenyi Wang, Li Shen, Yu Yao, Zhuo Huang, Shiming Chen, Chuanwu Yang, Mingming Gong, and Tongliang Liu. Improving non-transferable representation learning by harnessing content and style. In *The Twelfth International Conference on Learning Representations*, 2024.

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37, 2022a.

Li Hu, Jin Li, Guanbiao Lin, Shiyu Peng, Zhenxin Zhang, Yingying Zhang, and Changyu Dong. Defending against membership inference attacks with high utility by gan. *IEEE Transactions on Dependable and Secure Computing*, 20(3):2144–2157, 2022b.

Zixuan Hu, Li Shen, Zhenyi Wang, Tongliang Liu, Chun Yuan, and Dacheng Tao. Architecture, dataset and model-scale agnostic data-free meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7736–7745, 2023a.

Zixuan Hu, Li Shen, Zhenyi Wang, Baoyuan Wu, Chun Yuan, and Dacheng Tao. Learning to learn from apis: Black-box data-free meta-learning. In *International Conference on Machine Learning*, pp. 13610–13627. PMLR, 2023b.

Zixuan Hu, Li Shen, Zhenyi Wang, Yongxian Wei, and Dacheng Tao. Adaptive defense against harmful fine-tuning via bayesian data scheduler. In *Conference on Neural Information Processing Systems*, 2025.

Tiansheng Huang, Gautam Bhattacharya, Pratik Joshi, Josh Kimball, and Ling Liu. Antidote: Post-fine-tuning safety alignment for large language models against harmful fine-tuning. *arXiv preprint arXiv:2408.09600*, 2024a.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Tekin, and Ling Liu. Lisa: Lazy safety alignment for large language models against harmful fine-tuning attack. *Advances in Neural Information Processing Systems*, 37:104521–104555, 2024b.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. Booster: Tackling harmful fine-tuning for large language models via attenuating harmful perturbation. *arXiv preprint arXiv:2409.01586*, 2024c.

Tiansheng Huang, Sihao Hu, and Ling Liu. Vaccine: Perturbation-aware alignment for large language models against harmful fine-tuning attack. *Advances in Neural Information Processing Systems*, 37:74058–74088, 2024d.

Xiaowei Huang, Wenjie Ruan, Wei Huang, Gaojie Jin, Yi Dong, Changshun Wu, Saddek Bensalem, Ronghui Mu, Yi Qi, Xingyu Zhao, et al. A survey of safety and trustworthiness of large language models through the lens of verification and validation. *Artificial Intelligence Review*, 57(7):175, 2024e.

Dennis Jacob, Hend Alzahrani, Zhanhao Hu, Basel Alomair, and David Wagner. Promptshield: Deployable detection for prompt injection attacks. In *Proceedings of the Fifteenth ACM Conference on Data and Application Security and Privacy*, pp. 341–352, 2024.

Matthew Jagielski, Nicholas Carlini, David Berthelot, Alex Kurakin, and Nicolas Papernot. High accuracy and high fidelity extraction of neural networks. In *29th USENIX security symposium (USENIX Security 20)*, pp. 1345–1362, 2020.

Ismat Jarin and Birhanu Eshete. Miashield: Defending membership inference attacks via preemptive exclusion of members. *Proceedings on Privacy Enhancing Technologies*, 2023.

Hengrui Jia, Christopher A Choquette-Choo, Varun Chandrasekaran, and Nicolas Papernot. Entangled watermarks as a defense against model extraction. In *30th USENIX security symposium (USENIX Security 21)*, pp. 1937–1954, 2021.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International conference on machine learning*, pp. 2304–2313. PMLR, 2018.

Wenbo Jiang, Hongwei Li, Guowen Xu, Tianwei Zhang, and Rongxing Lu. A comprehensive defense framework against model extraction attacks. *IEEE Transactions on Dependable and Secure Computing*, 21(2): 685–700, 2023a.

Yujing Jiang, Xingjun Ma, Sarah Monazam Erfani, and James Bailey. Unlearnable examples for time series. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 213–225. Springer, 2024.

Yuxin Jiang, Chunkit Chan, Mingyang Chen, and Wei Wang. Lion: Adversarial distillation of proprietary large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3134–3154, 2023b.

Akshit Jindal, Vikram Goyal, Saket Anand, and Chetan Arora. Army of thieves: Enhancing black-box model extraction via ensemble based sample selection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3823–3832, 2024.

Mika Juuti, Sebastian Szyller, Samuel Marchal, and N Asokan. Prada: protecting against dnn model stealing attacks. In *2019 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 512–527. IEEE, 2019.

Mostafa Kahla, Si Chen, Hoang Anh Just, and Ruoxi Jia. Label-only model inversion attacks via boundary repulsion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15045–15053, 2022.

Sanjay Kariyappa and Moinuddin K Qureshi. Defending against model stealing attacks with adaptive misinformation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Sanjay Kariyappa, Atul Prakash, and Moinuddin K Qureshi. Maze: Data-free model stealing attack using zeroth-order gradient estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13814–13823, 2021a.

Sanjay Kariyappa, Atul Prakash, and Moinuddin K Qureshi. Protecting {dnn}s from theft using an ensemble of diverse models. In *International Conference on Learning Representations*, 2021b.

Pratik Karmakar and Debabrota Basu. Marich: A query-efficient distributionally equivalent model extraction attack. *Advances in Neural Information Processing Systems*, 36, 2024.

Manish Kesarwani, Bhaskar Mukhoty, Vijay Arya, and Sameep Mehta. Model extraction warning in mlaas paradigm. In *Proceedings of the 34th annual computer security applications conference*, pp. 371–380, 2018.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International Conference on Machine Learning*, pp. 17061–17084. PMLR, 2023.

Pang Wei Koh, Jacob Steinhardt, and Percy Liang. Stronger data poisoning attacks break data sanitization defenses. *Machine Learning*, 111(1):1–47, 2022.

Kalpesh Krishna, Gaurav Singh Tomar, Ankur P Parikh, Nicolas Papernot, and Mohit Iyyer. Thieves on sesame street! model extraction of bert-based apis. In *International Conference on Learning Representations*, 2020.

Keita Kurita, Paul Michel, and Graham Neubig. Weight poisoning attacks on pretrained models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2793–2806, 2020.

Ankita Laad and Khushboo Sawant. A literature review of various techniques to perform encryption and decryption of data. In *2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT)*, pp. 696–699. IEEE, 2021.

T Naga Lakshmi, S Jyothi, and M Rudra Kumar. Image encryption algorithms using machine learning and deep learning techniques—a survey. In *Modern Approaches in Machine Learning and Cognitive Science: A Walkthrough: Latest Trends in AI, Volume 2*, pp. 507–515. Springer, 2021.

Hongkyu Lee, Jeehyeong Kim, Seyoung Ahn, Rasheed Hussain, Sunghyun Cho, and Junggab Son. Digestive neural networks: A novel defense strategy against inference attacks in federated learning. *computers & security*, 109:102378, 2021.

Ang Li, Yin Zhou, Vethavikashini Chithrra Raghuram, Tom Goldstein, and Micah Goldblum. Commercial llm agents are already vulnerable to simple yet dangerous attacks. *arXiv preprint arXiv:2502.08586*, 2025a.

Chenxi Li, Abhinav Kumar, Zhen Guo, Jie Hou, and Reza Tourani. Unveiling the unseen: Exploring whitebox membership inference through the lens of explainability. *arXiv preprint arXiv:2407.01306*, 2024a.

Guanlin Li, Guowen Xu, Shangwei Guo, Han Qiu, Jiwei Li, and Tianwei Zhang. Extracting robust models with uncertain examples. In *The Eleventh International Conference on Learning Representations*, 2022.

Jiacheng Li, Ninghui Li, and Bruno Ribeiro. {MIST}: Defending against membership inference attacks through {Membership-Invariant} subspace training. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 2387–2404, 2024b.

Tong Li, Bingwen Feng, Guofeng Li, Xinzhen Li, Mingjin He, and Peiya Li. Visible watermark removal based on dual-input network. In *Proceedings of the 2021 ACM International Conference on Intelligent Computing and its Emerging Applications*, pp. 46–52, 2021a.

Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *International Conference on Learning Representations*.

Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Anti-backdoor learning: Training clean models on poisoned data. *Advances in Neural Information Processing Systems*, 34:14900–14912, 2021b.

Yuying Li, Gaoyang Liu, Chen Wang, and Yang Yang. Generating is believing: Membership inference attacks against retrieval-augmented generation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025b.

Zongjie Li, Chaozheng Wang, Pingchuan Ma, Chaowei Liu, Shuai Wang, Daoyuan Wu, Cuiyun Gao, and Yang Liu. On extracting specialized code abilities from large language models: A feasibility study. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, pp. 1–13, 2024c.

Jiacheng Liang, Ren Pang, Changjiang Li, and Ting Wang. Model extraction attacks revisited. In *Proceedings of the 19th ACM Asia Conference on Computer and Communications Security*, pp. 1231–1245, 2024.

Jing Liang, Li Niu, Fengjun Guo, Teng Long, and Liqing Zhang. Visible watermark removal via self-calibrated localization and background refinement. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 4426–4434, 2021.

Ching-Yung Lin and Shih-Fu Chang. Robust image authentication method surviving jpeg lossy compression. In *Storage and Retrieval for Image and Video Databases VI*, volume 3312, pp. 296–307. SPIE, 1997.

Xiao Lin, Zhining Liu, Dongqi Fu, Ruizhong Qiu, and Hanghang Tong. Backtime: Backdoor attacks on multivariate time series forecasting. *Advances in Neural Information Processing Systems*, 37:131344–131368, 2024.

Zijun Lin, Ke Xu, Chengfang Fang, Huadi Zheng, Aneez Ahmed Jaheezuddin, and Jie Shi. Quda: Query-limited data-free model extraction. In *Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security*, pp. 913–924, 2023.

Aiwei Liu, Leyi Pan, Yijian Lu, Jingjing Li, Xuming Hu, Xi Zhang, Lijie Wen, Irwin King, Hui Xiong, and Philip Yu. A survey of text watermarking in the era of large language models. *ACM Computing Surveys*, 57(2):1–36, 2024a.

Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Yu Cheng, and Wei Hu. A survey of attacks on large vision–language models: Resources, advances, and future trends. *IEEE Transactions on Neural Networks and Learning Systems*, 2025a.

Guanlin Liu and Lifeng Lai. Provably efficient black-box action poisoning attacks against reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12400–12410, 2021.

Guozhi Liu, Weiwei Lin, Qi Mu, Tiansheng Huang, Ruichao Mo, Yuren Tao, and Li Shen. Targeted vaccine: Safety alignment for large language models against harmful fine-tuning via layer-wise perturbation. *IEEE Transactions on Information Forensics and Security*, 2025b.

Han Liu, Yuhao Wu, Zhiyuan Yu, and Ning Zhang. Please tell me more: Privacy impact of explainability through the lens of membership inference attack. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 4791–4809. IEEE, 2024b.

Xiaogeng Liu, Minghui Li, Haoyu Wang, Shengshan Hu, Dengpan Ye, Hai Jin, Libing Wu, and Chaowei Xiao. Detecting backdoors during the inference stage based on corruption robustness consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16363–16372, 2023.

Yang Liu, Zhen Zhu, and Xiang Bai. Wdnet: Watermark-decomposition network for visible watermark removal. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 3685–3693, 2021.

Yepeng Liu, Yiren Song, Hai Ci, Yu Zhang, Haofan Wang, Mike Zheng Shou, and Yuheng Bu. Image watermarks are removable using controllable regeneration from clean noise. In *The Thirteenth International Conference on Learning Representations*.

Yixin Liu, Kaidi Xu, Xun Chen, and Lichao Sun. Stable unlearnable example: Enhancing the robustness of unlearnable examples via stable error-minimizing noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 3783–3791, 2024c.

Siyu Luan, Zhenyi Wang, Li Shen, Zonghua Gu, Chao Wu, and Dacheng Tao. Dynamic neural fortresses: An adaptive shield for model extraction defense. In *The Thirteenth International Conference on Learning Representations*.

Shweta Mahajan, Tanzila Rahman, Kwang Moo Yi, and Leonid Sigal. Prompting hard or hardly prompting: Prompt inversion for text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6808–6817, 2024.

Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5188–5196, 2015.

Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. Llm dataset inference: Did you train on my dataset? *Advances in Neural Information Processing Systems*, 37:124069–124092, 2024.

Disha Makhija, Manoj Ghuhan Arivazhagan, Vinayshekhar Bannihatti Kumar, and Rashmi Gangadharaiah. Neural breadcrumbs: Membership inference attacks on llms through hidden state and attention pattern analysis. *arXiv preprint arXiv:2509.05449*, 2025.

VM Manikandan and V Masilamani. Reversible data hiding scheme during encryption using machine learning. *Procedia computer science*, 133:348–356, 2018.

Yanxu Mao, Tiehan Cui, Peipei Liu, Datao You, and Hongsong Zhu. From llms to mllms to agents: A survey of emerging paradigms in jailbreak attacks and defenses within llm ecosystem. *arXiv preprint arXiv:2506.15170*, 2025.

AprilPyone MaungMaung and Hitoshi Kiya. Generative model-based attack on learnable image encryption for privacy-preserving deep learning. *arXiv preprint arXiv:2303.05036*, 2023.

Ruohan Meng, Chenyu Yi, Yi Yu, Siyuan Yang, Bingquan Shen, and Alex C Kot. Semantic deep hiding for robust unlearnable examples. *IEEE Transactions on Information Forensics and Security*, 19:6545–6558, 2024.

Yibo Miao, Yifan Zhu, Lijia Yu, Jun Zhu, Xiao-Shan Gao, and Yinpeng Dong. T2vsafetybench: Evaluating the safety of text-to-video generative models. *Advances in Neural Information Processing Systems*, 37: 63858–63872, 2024.

Yash More, Prakhar Ganesh, and Golnoosh Farnadi. Towards more realistic extraction attacks: An adversarial perspective. *CoRR*, 2024.

John X Morris, Wenting Zhao, Justin T Chiu, Vitaly Shmatikov, and Alexander M Rush. Language model inversion. *arXiv preprint arXiv:2311.13647*, 2023.

Seyed Mojtaba Mousavi, Alireza Naghsh, and SAR Abu-Bakar. Watermarking techniques used in medical images: a survey. *Journal of digital imaging*, 27(6):714–729, 2014.

Sumit Mukherjee, Yixi Xu, Anusua Trivedi, Nabajyoti Patowary, and Juan L Ferres. privgan: Protecting gans from membership inference attacks at low cost to utility. *Proceedings on Privacy Enhancing Technologies*, 2021.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, pp. 634–646, 2018.

Milad Nasr, Javier Rando, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from aligned, production language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Bao-Ngoc Nguyen, Keshigeyan Chandrasegaran, Milad Abdollahzadeh, and Ngai-Man Man Cheung. Label-only model inversion attacks via knowledge transfer. *Advances in Neural Information Processing Systems*, 36, 2024a.

Thanh Toan Nguyen, Nguyen Quoc Viet Hung, Thanh Tam Nguyen, Thanh Trung Huynh, Thanh Thi Nguyen, Matthias Weidlich, and Hongzhi Yin. Manipulating recommender systems: A survey of poisoning attacks and countermeasures. *ACM Computing Surveys*, 57(1):1–39, 2024b.

Li Niu, Xing Zhao, Bo Zhang, and Liqing Zhang. Fine-grained visible watermark removal. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12770–12779, 2023.

Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Knockoff nets: Stealing functionality of black-box models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4954–4963, 2019.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Soham Pal, Yash Gupta, Aditya Shukla, Aditya Kanade, Shirish Shevade, and Vinod Ganapathy. Activethief: Model extraction using active learning and unannotated public data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 865–872, 2020.

Leyi Pan, Aiwei Liu, Shiyu Huang, Yijian Lu, Xuming Hu, Lijie Wen, Irwin King, and Philip S Yu. Can llm watermarks robustly prevent unauthorized knowledge distillation? *CoRR*, 2025.

Michael-Andrei Panaitescu-Liess, Zora Che, Bang An, Yuancheng Xu, Pankayaraj Pathmanathan, Souradip Chakraborty, Sicheng Zhu, Tom Goldstein, and Furong Huang. Can watermarking large language models prevent copyrighted text generation and hide training data? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25002–25009, 2025.

Yan Pang, Tianhao Wang, Xuhui Kang, Mengdi Huai, and Yang Zhang. White-box membership inference attacks against diffusion models. *arXiv preprint arXiv:2308.06405*, 2023.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pp. 506–519, 2017.

Sen Peng, Yufei Chen, Cong Wang, and Xiaohua Jia. Protecting the intellectual property of diffusion models by the watermark diffusion process. *CoRR*, 2023.

Xiong Peng, Feng Liu, Jingfeng Zhang, Long Lan, Junjie Ye, Tongliang Liu, and Bo Han. Bilateral dependency optimization: Defending against model-inversion attacks. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 1358–1367, 2022.

Yuefeng Peng, Jaechul Roh, Subhransu Maji, and Amir Houmansadr. Oslo: one-shot label-only membership inference attacks. *Advances in Neural Information Processing Systems*, 37:62310–62333, 2024.

Julien Piet, Chawin Sitawarin, Vivian Fang, Norman Mu, and David Wagner. Markmywords: Analyzing and evaluating language model watermarks. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 68–91. IEEE, 2025.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *ICLR*, 2024.

Miguel A Ramirez, Song-Kyoo Kim, Hussam Al Hamadi, Ernesto Damiani, Young-Ji Byon, Tae-Yeon Kim, Chung-Suk Cho, and Chan Yeob Yeun. Poisoning attacks and defenses on artificial intelligence: A survey. *arXiv preprint arXiv:2202.10276*, 2022.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International conference on machine learning*, pp. 4334–4343. PMLR, 2018.

Nuria Rodríguez-Barroso, Daniel Jiménez-López, M Victoria Luzón, Francisco Herrera, and Eugenio Martínez-Cámara. Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges. *Information Fusion*, 90:148–173, 2023.

David Rolnick and Konrad Kording. Reverse-engineering deep relu networks. In *International conference on machine learning*, pp. 8178–8187. PMLR, 2020.

Domenic Rosati, Jan Wehner, Kai Williams, Łukasz Bartoszcze, Jan Batzner, Hassan Sajjad, and Frank Rudzicz. Immunization against harmful fine-tuning attacks. *arXiv preprint arXiv:2402.16382*, 2024a.

Domenic Rosati, Jan Wehner, Kai Williams, Lukasz Bartoszcze, Robie Gonzales, Subhabrata Majumdar, Hassan Sajjad, Frank Rudzicz, et al. Representation noising: A defence mechanism against harmful finetuning. *Advances in Neural Information Processing Systems*, 37:12636–12676, 2024b.

Jonathan Rosenthal, Eric Enouen, Hung Viet Pham, and Lin Tan. Disguide: Disagreement-guided data-free model extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9614–9622, 2023.

Vinu Sankar Sadasivan, Mahdi Soltanolkotabi, and Soheil Feizi. Cuda: Convolution-based unlearnable datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3862–3871, 2023.

Amir Mahdi Sadeghzadeh, Amir Mohammad Sobhanian, Faezeh Dehghan, and Rasool Jalili. Hoda: Hardness-oriented detection of model extraction attacks. *IEEE Transactions on Information Forensics and Security*, 19:1429–1439, 2023.

Sunandini Sanyal, Sravanti Addepalli, and R Venkatesh Babu. Towards data-free model stealing in a hard label setting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15284–15293, 2022.

Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. *Advances in neural information processing systems*, 31, 2018.

Mingwen Shao, Lingzhuang Meng, Yuanjian Qiao, Lixu Zhang, and Wangmeng Zuo. Latent code augmentation based on stable diffusion for data-free substitute attacks. *IEEE Transactions on Neural Networks and Learning Systems*, 2025a.

Shuai Shao, Qihan Ren, Chen Qian, Boyi Wei, Dadi Guo, Jingyi Yang, Xinhao Song, Linfeng Zhang, Weinan Zhang, Dongrui Liu, et al. Your agent may misevolve: Emergent risks in self-evolving llm agents. *arXiv preprint arXiv:2509.26354*, 2025b.

Virat Shejwalkar and Amir Houmansadr. Membership privacy for machine learning models through knowledge transfer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 9549–9557, 2021.

Yanyao Shen and Sujay Sanghavi. Learning with bad training data via iterative trimmed loss minimization. In *International conference on machine learning*, pp. 5739–5748. PMLR, 2019.

Haonan Shi, Tu Ouyang, and An Wang. Learning-based difficulty calibration for enhanced membership inference attacks. In *2024 IEEE 9th European Symposium on Security and Privacy (EuroS&P)*, pp. 62–77. IEEE, 2024.

Warit Sirichotedumrong and Hitoshi Kiya. Visual security evaluation of learnable image encryption methods against ciphertext-only attacks. In *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1304–1309. IEEE, 2020.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1): 1929–1958, 2014.

Peixian Su and Yong Zhang. Deep learning for visible watermark removal: A survey. *Computational Intelligence*, 41(3):e70072, 2025.

Nagesh Subbanna, Matthias Wilms, Anup Tuladhar, and Nils D Forkert. An analysis of the vulnerability of two common deep learning-based medical image segmentation techniques to model inversion attacks. *Sensors*, 21(11):3874, 2021.

Ruizhou Sun, Yukun Su, and Qingyao Wu. Denet: Disentangled embedding network for visible watermark removal. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2411–2419, 2023.

Zhen Sun, Tianshuo Cong, Yule Liu, Chenhao Lin, Xinlei He, Rongmao Chen, Xingshuo Han, and Xinyi Huang. Peftguard: detecting backdoor attacks against parameter-efficient fine-tuning. In *2025 IEEE Symposium on Security and Privacy (SP)*, pp. 1713–1731. IEEE, 2025.

Anshuman Suri, Xiao Zhang, and David Evans. Do parameters reveal more than loss for membership inference? *arXiv preprint arXiv:2406.11544*, 2024.

Sebastian Szyller, Buse Gul Atli, Samuel Marchal, and N Asokan. Dawn: Dynamic adversarial watermarking of neural networks. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 4417–4425, 2021.

Tatsuya Takemura, Naoto Yanai, and Toru Fujiwara. Model extraction attacks on recurrent neural networks. *Journal of Information Processing*, 28:1010–1024, 2020.

Mingtian Tan, Xiaofei Xie, Jun Sun, and Tianhao Wang. Mitigating membership inference attacks via weighted smoothing. In *Proceedings of the 39th Annual Computer Security Applications Conference*, pp. 787–798, 2023.

Fengyi Tang, Wei Wu, Jian Liu, Huimei Wang, and Ming Xian. Privacy-preserving distributed deep learning via homomorphic re-encryption. *Electronics*, 8(4):411, 2019.

Minxue Tang, Anna Dai, Louis DiValentin, Aolin Ding, Amin Hass, Neil Zhenqiang Gong, Yiran Chen, et al. {ModelGuard}:{Information-Theoretic} defense against model extraction attacks. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 5305–5322, 2024.

Xinyu Tang, Saeed Mahloujifar, Liwei Song, Virat Shejwalkar, Milad Nasr, Amir Houmansadr, and Prateek Mittal. Mitigating membership inference attacks by {Self-Distillation} through a novel ensemble architecture. In *31st USENIX security symposium (USENIX security 22)*, pp. 1433–1450, 2022.

Shuchang Tao, Qi Cao, Huawei Shen, Junjie Huang, Yunfan Wu, and Xueqi Cheng. Single node injection attack against graph neural networks. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 1794–1803, 2021.

Zhiyi Tian, Lei Cui, Jie Liang, and Shui Yu. A comprehensive survey on poisoning attacks and countermeasures in machine learning. *ACM Computing Surveys*, 55(8):1–35, 2022.

Antonios Tragoudaras, Theofanis Aslanidis, Emmanouil Georgios Lionis, Marina Orozco González, and Panagiotis Eustratiadis. Information leakage of sentence embeddings via generative embedding inversion attacks. *arXiv preprint arXiv:2504.16609*, 2025.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction {APIs}. In *25th USENIX security symposium (USENIX Security 16)*, pp. 601–618, 2016.

Toan Tran, Ruixuan Liu, and Li Xiong. Tokens for learning, tokens for unlearning: Mitigating membership inference attacks in large language models via dual-purpose training. *arXiv preprint arXiv:2502.19726*, 2025.

Jean-Baptiste Truong, Pratyush Maini, Robert J Walls, and Nicolas Papernot. Data-free model extraction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4771–4780, 2021.

Eric Wallace, Olivia Watkins, Miles Wang, Kai Chen, and Chris Koch. Estimating worst-case frontier risks of open-weight llms. *arXiv preprint arXiv:2508.03153*, 2025.

Wenbo Wan, Jun Wang, Yunming Zhang, Jing Li, Hui Yu, and Jiande Sun. A comprehensive survey on robust image watermarking. *Neurocomputing*, 488:226–247, 2022.

Kai Wang, Jinxia Wu, Tianqing Zhu, Wei Ren, and Ying Hong. Defense against membership inference attack in graph neural networks through graph perturbation. *International Journal of Information Security*, 22 (2):497–509, 2023a.

Kuan-Chieh Wang, Yan Fu, Ke Li, Ashish Khisti, Richard Zemel, and Alireza Makhzani. Variational model inversion attacks. *Advances in Neural Information Processing Systems*, 34:9706–9719, 2021a.

Xiuling Wang and Wendy Hui Wang. Gcl-leak: Link membership inference attacks against graph contrastive learning. *Proceedings on Privacy Enhancing Technologies*, 2024.

Yijue Wang, Chenghong Wang, Zigeng Wang, Shanglin Zhou, Hang Liu, Jinbo Bi, Caiwen Ding, and Sanguthevar Rajasekaran. Against membership inference attack: Pruning is all you need. In *30th International Joint Conference on Artificial Intelligence, IJCAI 2021*, pp. 3141–3147. International Joint Conferences on Artificial Intelligence, 2021b.

Zhenyi Wang, Li Shen, Tongliang Liu, Tiehang Duan, Yanjun Zhu, Donglin Zhan, David Doermann, and Mingchen Gao. Defending against data-free model extraction by distributionally robust defensive training. *Advances in Neural Information Processing Systems*, 36, 2023b.

Zhenyi Wang, Li Shen, Junfeng Guo, Tiehang Duan, Siyu Luan, Tongliang Liu, and Mingchen Gao. Training a secure model against data-free model extraction. In *European Conference on Computer Vision*, pp. 323–340. Springer, 2024a.

Zhenyi Wang, Yihan Wu, and Heng Huang. Defense against model extraction attack by bayesian active watermarking. In *Forty-first International Conference on Machine Learning*, 2024b.

Yongxian Wei, Zixuan Hu, Li Shen, Zhenyi Wang, Yu Li, Chun Yuan, and Dacheng Tao. Task groupings regularization: Data-free meta-learning with heterogeneous pre-trained models. In *International Conference on Machine Learning*, pp. 52573–52587. PMLR, 2024a.

Yongxian Wei, Zixuan Hu, Zhenyi Wang, Li Shen, Chun Yuan, and Dacheng Tao. Free: Faster and better data-free meta-learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23273–23282, 2024b.

Yongxian Wei, Zixuan Hu, Li Shen, Zhenyi Wang, Chun Yuan, and Dacheng Tao. Open-vocabulary customization from CLIP via data-free knowledge distillation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jing Wen, Siu-Ming Yiu, and Lucas CK Hui. Defending against model inversion attack by adversarial examples. In *2021 IEEE International Conference on Cyber Security and Resilience (CSR)*, pp. 551–556. IEEE, 2021.

Yuxin Wen, Yuchen Liu, Chen Chen, and Lingjuan Lyu. Detecting, explaining, and mitigating memorization in diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.

Dong-Dong Wu, Chilin Fu, Weichang Wu, Wenwen Xia, Xiaolu Zhang, Jun Zhou, and Min-Ling Zhang. Efficient model stealing defense with noise transition matrix. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24305–24315, 2024.

Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 3905–3911, 2018a.

Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. In *International Conference on Learning Representations*, 2018b.

Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. Dba: Distributed backdoor attacks against federated learning. In *International conference on learning representations*.

Yi Xie, Jie Zhang, Shiqian Zhao, Tianwei Zhang, and Xiaofeng Chen. Same: sample reconstruction against model extraction attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19974–19982, 2024.

Han Xu, Yao Ma, Hao-Chen Liu, Debayan Deb, Hui Liu, Ji-Liang Tang, and Anil K Jain. Adversarial attacks and defenses in images, graphs and text: A review. *International journal of automation and computing*, 17(2):151–178, 2020.

Huan Xu, Zhanhao Zhang, Xiaodong Yu, Yingbo Wu, Zhiyong Zha, Bo Xu, Wenfeng Xu, Menglan Hu, and Kai Peng. Targeted training data extraction—neighborhood comparison-based membership inference attacks in large language models. *Applied Sciences*, 14(16):7118, 2024a.

Nuo Xu, Binghui Wang, Ran Ran, Wujie Wen, and Parv Venkitasubramaniam. Neuguard: Lightweight neuron-guided defense against membership inference attacks. In *Proceedings of the 38th annual computer security applications conference*, pp. 669–683, 2022.

Yue Xu, Xiuyuan Qi, Zhan Qin, and Wenjie Wang. Cross-modality information check for detecting jailbreaking in multimodal large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13715–13726, 2024b.

Anli Yan, Teng Huang, Lishan Ke, Xiaozhang Liu, Qi Chen, and Changyu Dong. Explanation leaks: explanation-guided model extraction attacks. *Information Sciences*, 632:269–284, 2023.

Enneng Yang, Zhenyi Wang, Li Shen, Nan Yin, Tongliang Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. Continual learning from a stream of apis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024a.

Ruikang Yang, Jianfeng Ma, Yinbin Miao, and Xindi Ma. Privacy-preserving generative framework for images against membership inference attacks. *IET Communications*, 17(1):45–62, 2023.

Wencheng Yang, Song Wang, Di Wu, Taotao Cai, Yanming Zhu, Shicheng Wei, Yiying Zhang, Xu Yang, Zhaohui Tang, and Yan Li. Deep learning model inversion attacks and defenses: a comprehensive survey. *Artificial Intelligence Review*, 58(8):242, 2025a.

Yunfei Yang, Xiaojun Chen, Yuexin Xuan, and Zhendong Zhao. Dualcos: Query-efficient data-free model stealing with dual clone networks and optimal samples. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2024b.

Zhiguang Yang, Gejian Zhao, and Hanzhou Wu. Watermarking for large language models: A survey. *Mathematics*, 13(9):1420, 2025b.

Jingwei Yi, Rui Ye, Qisi Chen, Bin Zhu, Siheng Chen, Defu Lian, Guangzhong Sun, Xing Xie, and Fangzhao Wu. On the vulnerability of safety alignment in open-access llms. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 9236–9260, 2024.

Qichao Ying, Hang Zhou, Zhenxing Qian, Sheng Li, and Xinpeng Zhang. Learning to immunize images for tamper localization and self-recovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13814–13830, 2023.

Honggang Yu, Kaichen Yang, Teng Zhang, Yun-Yun Tsai, Tsung-Yi Ho, and Yier Jin. Cloudleak: Large-scale deep learning models stealing through adversarial examples. In *NDSS*, volume 6, pp. 3, 2020.

Xiaojian Yuan, Kejiang Chen, Wen Huang, Jie Zhang, Weiming Zhang, and Nenghai Yu. Data-free hard-label robustness stealing attack. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6853–6861, 2024.

Xiaoyong Yuan, Leah Ding, Lan Zhang, Xiaolin Li, and Dapeng Oliver Wu. Es attack: Model stealing against deep neural networks without data hurdles. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(5):1258–1270, 2022.

Zhenrui Yue, Zhankui He, Huimin Zeng, and Julian McAuley. Black-box attacks on sequential recommenders via data-free model extraction. In *Proceedings of the 15th ACM conference on recommender systems*, pp. 44–54, 2021.

Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-cost high-power membership inference attacks. *arXiv preprint arXiv:2312.03262*, 2023.

Yi Zeng, Yu Yang, Andy Zhou, Jeffrey Ziwei Tan, Yuheng Tu, Yifan Mai, Kevin Klyman, Minzhou Pan, Ruoxi Jia, Dawn Song, et al. Air-bench 2024: A safety benchmark based on regulation and policies specified risk categories. In *The Thirteenth International Conference on Learning Representations*.

Chuan Zhang, Haotian Liang, Zhuopeng Li, Tong Wu, Licheng Wang, and Liehuang Zhu. Ptbstolen: Pre-trained encoder stealing through perturbed samples. In *International Symposium on Emerging Information Security and Applications*, pp. 1–19. Springer, 2023a.

Haitian Zhang, Guang Hua, Xinya Wang, Hao Jiang, and Wen Yang. Categorical inference poisoning: verifiable defense against black-box dnn model stealing without constraining surrogate data and query times. *IEEE Transactions on Information Forensics and Security*, 18:1473–1486, 2023b.

Haitian Zhang, Guang Hua, and Wen Yang. Poisoning-free defense against black-box model extraction. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4760–4764. IEEE, 2024.

Jialong Zhang, Zhongshu Gu, Jiyong Jang, Hui Wu, Marc Ph Stoecklin, Heqing Huang, and Ian Molloy. Protecting intellectual property of deep neural networks with watermarking. In *Proceedings of the 2018 on Asia conference on computer and communications security*, pp. 159–172, 2018.

Jie Zhang, Chen Chen, and Lingjuan Lyu. Ideal: Query-efficient data-free learning from black-box models. In *The Eleventh International Conference on Learning Representations*, 2022.

Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. The secret revealer: Generative model-inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 253–261, 2020.

Zhanyuan Zhang, Yizheng Chen, and David Wagner. Seat: Similarity encoder by adversarial training for detecting model extraction attack queries. In *Proceedings of the 14th ACM Workshop on artificial intelligence and security*, pp. 37–48, 2021.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.

Xing Zhao, Li Niu, and Liqing Zhang. Visible watermark removal with dynamic kernel and semantic-aware propagation. In *BMVC*, volume 1, pp. 3, 2022.

Xuandong Zhao, Kexun Zhang, Zihao Su, Saastha Vasan, Ilya Grishchenko, Christopher Kruegel, Giovanni Vigna, Yu-Xiang Wang, and Lei Li. Invisible image watermarks are provably removable using generative ai. *Advances in neural information processing systems*, 37:8643–8672, 2024a.

Xuejun Zhao, Wencan Zhang, Xiaokui Xiao, and Brian Lim. Exploiting explanations for model inversion attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 682–692, 2021.

Yunlong Zhao, Xiaoheng Deng, Yijing Liu, Xinjun Pei, Jiazhi Xia, and Wei Chen. Fully exploiting every real sample: Superpixel sample gradient model stealing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24316–24325, 2024b.

Zhengyue Zhao, Xiaogeng Liu, Somesh Jha, Patrick McDaniel, Bo Li, and Chaowei Xiao. Can watermarks be used to detect llm ip infringement for free? In *The Thirteenth International Conference on Learning Representations*.

Junxiang Zheng, Yongzhi Cao, and Hanpin Wang. Resisting membership inference attacks through knowledge distillation. *Neurocomputing*, 452:114–126, 2021.

Da Zhong, Xiuling Wang, Zhichao Xu, Jun Xu, and Wendy Hui Wang. Interaction-level membership inference attack against recommender systems with long-tailed distribution. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 3433–3442, 2024.

Baohang Zhou, Zezhong Wang, Lingzhi Wang, Hongru Wang, Ying Zhang, Kehui Song, Xuhui Sui, and Kam-Fai Wong. Dpdllm: A black-box framework for detecting pre-training data from large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 644–653, 2024.

Zhanke Zhou, Chenyu Zhou, Xuan Li, Jiangchao Yao, Quanming Yao, and Bo Han. On strengthening and defending graph reconstruction attack with markov chain approximation. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 42843–42877, 2023.

Chen Zhu, W Ronny Huang, Hengduo Li, Gavin Taylor, Christoph Studer, and Tom Goldstein. Transferable clean-label poisoning attacks on deep neural nets. In *International conference on machine learning*, pp. 7614–7623. PMLR, 2019.

Gongxi Zhu, Donghao Li, Hanlin Gu, Yuxing Han, Yuan Yao, Lixin Fan, and Qiang Yang. Evaluating membership inference attacks and defenses in federated learning. *CoRR*, 2024a.

Hongyu Zhu, Sichu Liang, Wentao Hu, Li Fangqi, Ju Jia, and Shi-Lin Wang. Reliable model watermarking: Defending against theft without compromising on evasion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10124–10133, 2024b.

Hongyu Zhu, Wentao Hu, Sichu Liang, Fangqi Li, Wenwen Wang, and Shilin Wang. Efficient and effective model extraction. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025a.

Meiyi Zhu, Caili Guo, Chunyan Feng, and Osvaldo Simeone. On the impact of uncertainty and calibration on likelihood-ratio membership inference attacks. *IEEE Transactions on Information Forensics and Security*, 2025b.

Zihang Zou, Boqing Gong, and Liqiang Wang. Anti-neuron watermarking: Protecting personal data against unauthorized neural networks. In *European Conference on Computer Vision*, pp. 449–465. Springer, 2022.

Zihang Zou, Boqing Gong, and Liqiang Wang. Attention to neural plagiarism: Diffusion models can plagiarize your copyrighted images! In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19546–19556, 2025.