Adversarial Dynamics in Generative AI

Abir Bokhtiar

***Abstract*—The rapid proliferation of Foundation Models, encompassing Large Language Models (LLMs) and Diffusion Models, has introduced a complex security landscape distinct from traditional discriminative AI. While reinforcement learning strategies have improved model alignment, recent literature demonstrates that generative models remain brittle against adversarial perturbations, prompting a continuous arms race between attack generation and defense mechanisms. This systematic review, conducted in accordance with PRISMA 2020 guidelines, synthesizes 76 primary studies to construct a rigorous taxonomy of adversarial threats targeting Generative AI. This review explicitly differentiates between *safety* (alignment failures) and *robustness* (adversarial susceptibility), analyzing high-impact vectors including universal jailbreaks, prompt injection, and multimodal backdoors. Furthermore, this study critically evaluates the efficacy of current defenses—from certified robustness to unlearning—highlighting significant theoretical gaps in securing multimodal agents.**

I. Introduction

The transition from discriminative to generative artificial intelligence has fundamentally altered the threat landscape of machine learning [1]. Foundation models [2], particularly Large Language Models (LLMs) and Text-to-Image Diffusion Models [3], have achieved unprecedented capabilities in content synthesis. However, this utility is coupled with severe security risks. As these models are integrated into agentic systems [4] and real-world applications, they expose novel attack surfaces ranging from data leakage [5] to the generation of harmful content via "jailbreaking" [6].

A critical distinction in this domain lies between *safety* and *robustness*, concepts often conflated in early literature. Safety generally refers to a model's alignment with human values and its refusal to generate harmful content under standard conditions [7]. Robustness, conversely, measures a model's resistance to worst-case adversarial inputs designed to force failure [1]. Recent studies indicate that while safety training techniques like Reinforcement Learning from Human Feedback (RLHF) [8] reduce casual misuse, they often fail to provide robust guarantees against optimized adversarial attacks. For instance, Wei et al. [9] demonstrate that safety training can be circumvented due to conflicting training objectives, while Hubinger et al. [10] reveal that deceptive alignment can persist even after rigorous safety tuning.

The complexity of these threats is magnified in multimodal contexts. Attacks are no longer confined to semantic manipulation of text; visual adversarial examples can now "hijack" the behavior of multimodal systems [11], and visual prompts can bypass text-based safety filters in Vision-Language Models (VLMs) [12], [13]. Furthermore, the democratization of powerful open-weights models [14] has accelerated the development of automated attack frameworks, such as Genetic Coordinate Gradient (GCG) [6] and Tree of Attacks (TAP) [15], which lower the barrier to entry for adversaries.

While several recent surveys have mapped the broad landscape of Generative AI security [16]–[18], they often prioritize breadth over a mechanism-focused analysis of the attack-defense dynamic. Existing reviews frequently lack a rigorous separation of white-box versus black-box realism and insufficiently address the emerging risks in agentic workflows [19]. Moreover, the rapid evolution of "indirect prompt injection" [20] and supply-chain poisoning [21] necessitates an updated, systematic categorization of threats that extends beyond simple evasion.

A. Contributions

This paper addresses these gaps via a PRISMA-compliant systematic review. The major contributions are:

- **Rigorous Taxonomy of Attacks:** This study classifies threats not merely by modality, but by the interaction surface (e.g., prompt injection vs. latent space perturbation) and attacker capability (zero-query transferability vs. adaptive optimization), synthesizing key findings from seminal works like [5], [6], [22].
- **Safety vs. Robustness Delineation:** The comprehensive analysis in this study outlines a clear separation between compliance-based safety failures (e.g., [23]) and adversarial robustness failures (e.g., [24]), arguing that current alignment techniques are insufficient for the latter.
- **Multimodal Integration:** This work analyzes the unique vulnerabilities introduced by late-fusion architectures in VLMs, where visual inputs can act as Trojan triggers [25] or bypass textual guardrails [26].
- **Critical Defense Evaluation:** The study systematically evaluates defense categories, contrasting the theoretical guarantees of randomized smoothing [24], [27] against the practical limitations of input filtering [28] and unlearning [29], [30].

B. Paper Organization

The remainder of this paper is organized as follows: Section II establishes the theoretical background and threat models. Section III details the PRISMA methodology used for study selection. Section IV and V present the taxonomies of attacks and defenses, respectively. Section VI offers a comparative analysis of trends, followed by evaluation metrics in Section VII. Section VIII discusses open challenges, and Section IX concludes.

II. Background and Preliminaries

Generative AI systems can be viewed as probabilistic function approximators that learn to model high-dimensional

TABLE I
LIST OF ACRONYMS AND ABBREVIATIONS

Acronym	Definition
	Models & Architectures
LLM	Large Language Model
VLM	Vision-Language Model
LDM	Latent Diffusion Model
ViT	Vision Transformer
RAG	Retrieval-Augmented Generation
	Attacks & Exploits
GCG	Greedy Coordinate Gradient (Optimization Attack)
PAIR	Prompt Automatic Iterative Refinement
TAP	Tree of Attacks with Pruning
ASR	Attack Success Rate
	Defenses & Alignment
RLHF	Reinforcement Learning from Human Feedback
DPO	Direct Preference Optimization
HALO	Human Preference Aligned Offline Reward Learning
AT	Adversarial Training

data distributions. At deployment time, these models generate outputs by sampling from learned distributions conditioned on user inputs. From a security perspective, this means that any controllable input channel like text, image, memory, or tool interaction can become an attack surface. Therefore, understanding the internal generation pipeline is critical to understanding how adversarial manipulation propagates through the system. Furthermore, to systematically analyze vulnerabilities in Generative AI, we must first define the architectural paradigms of the victim models and formalize the adversary's operational capabilities. This section distinguishes between the underlying generation mechanisms and the alignment processes designed to constrain them.

A. Generative Model Architectures

1) Large Language Models (LLMs): Modern LLMs function as autoregressive probability distributions over sequences of tokens. Given a vocabulary $\mathcal{V}$ and a context window of tokens $x_{1:t}$, the model is trained to minimize the negative log-likelihood of the next token x_{t+1}:

$$\mathcal{L}_{LLM} = -\sum_{t=1}^{T} \log P_\theta(x_{t+1} \mid x_{1:t}) \tag{1}$$

where $P_\theta(x_{t+1} \mid x_{1:t}) = \text{softmax}(W_o h_t)$ and $h_t = \text{Transformer}(x_{1:t})$ is the hidden representation. While the architectural backbone is primarily the Transformer, the security surface is defined by the inference process. As noted by Huang et al. [31], the decoding strategy (e.g., temperature, top-k) introduces stochasticity that adversaries can exploit to bypass deterministic safety filters. Furthermore, the integration of these models into agentic workflows [4] expands the input space from pure text to tool-use directives, creating new vectors for indirect injection.

2) Latent Diffusion Models (LDMs): Text-to-Image (T2I) generation predominantly relies on diffusion processes, specifically Latent Diffusion Models (LDMs) as formalized by Rombach et al. [3]. Unlike pixel-space probabilistic models, LDMs operate in a compressed latent space $\mathcal{Z}$. The training objective involves predicting noise ϵ added to a latent representation z_t conditioned on a text prompt y:

$$\mathcal{L}_{LDM} = \mathbb{E}_{z,\epsilon,y,t}\left[||\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))||_2^2\right] \tag{2}$$

where z_t is the noisy latent representation, τ_θ is a text encoder (e.g., CLIP) and $\tau_\theta(y)$ encodes text conditioning (e.g., CLIP embedding). Security research highlights that the disjoint nature of the text encoder and the UNet denoiser creates distinct attack surfaces; for instance, Struppek et al. [32] demonstrate that the text encoder can be poisoned independently of the visual generator.

Because conditioning and generation are handled by separate modules (text encoder + UNet denoiser), attacks can target individual components, including encoder poisoning or adversarial conditioning manipulation [3].

3) Multimodal and Vision-Language Models (VLMs): VLMs, such as LLaVA or GPT-4V, typically employ a "late fusion" strategy where visual features from an encoder (e.g., ViT) are projected into the LLM's embedding space [12]. Formally, if $v = \text{ViT}(I)$ is the visual embedding of an image I, and x is a text token sequence, the multimodal representation is

$$h_t^{VLM} = h_t^{LLM} + W_v v \tag{3}$$

where W_v is a learned projection. This integration creates a continuous optimization path from pixel input to text output, allowing adversaries to perform gradient-based optimization on images to coerce specific textual behaviors (e.g., jailbreaks), as shown in [11], [25].

B. Alignment and Safety Training

Raw foundation models are rarely deployed directly; they undergo alignment to mitigate harmful outputs.

- **Reinforcement Learning from Human Feedback (RLHF):** Established by Christiano et al. [8] and refined for safety by Bai et al. [7], RLHF optimizes a reward model trained on human preferences. If r_ϕ is the reward model and π_θ the policy:

  $$\mathcal{L}_{RLHF} = -\mathbb{E}_{x\sim D}\left[r_\phi(f_\theta(x))\right] \tag{4}$$

  RLHF aligns models with human preferences through three stages:

  1) Supervised fine-tuning on human demonstrations
  2) Reward model training using human preference rankings
  3) Policy optimization (typically PPO) to maximize reward

  RLHF significantly improves helpfulness and harmlessness but is complex, requiring multiple models and reinforcement learning optimization [33], [34]. Studies like Wei et al. [9] argue that the competing objectives of "helpfulness" and "harmlessness" create optimization landscapes where safety can be compromised by highly "helpful" responses to malicious queries.

- **Direct Preference Optimization (DPO):** Rafailov et al. [34] proposed DPO to optimize the policy directly against preference data without an explicit reward model. The optimization objective is:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\sigma\Big(\beta\log\frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta\log\frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\Big)\Big] \tag{5}$$

  The model being optimized is denoted by $\pi_\theta(y \mid x)$, while $\pi_{\text{ref}}(y \mid x)$ represents the frozen reference policy. The training data consists of triplets (x, y_w, y_l) sampled from a preference dataset $\mathcal{D}$, where y_w and y_l are the preferred and dispreferred completions for prompt x, respectively. The hyperparameter β serves as a temperature coefficient controlling the deviation from the reference policy. Finally, σ denotes the logistic sigmoid function, and the loss is calculated as the expected value $\mathbb{E}$ over the dataset distributions. Standard DPO may under-penalize hazards unless explicitly weighted (e.g., HALO) [35].

Despite these measures, Wolf et al. [36] provide theoretical evidence suggesting that perfect alignment is impossible in high-dimensional spaces, implying that adversarial examples are an inevitable feature of current deep learning paradigms rather than a curable bug.

C. Threat Model and Attacker Capabilities

We adopt a formal threat model consistent with the taxonomies presented in [5], [6]:

$$\max_{\delta\in\Delta} \mathcal{L}(f_\theta(x \oplus \delta), y_{target}) \quad \text{s.t.} \quad P(x \oplus \delta) < \tau \tag{6}$$

where x is the clean prompt, δ is the adversarial perturbation (suffix or continuous noise), $\oplus$ denotes the injection operation, y_{target} is the prohibited behavior, and $P(\cdot)$ represents a perplexity or detectability filter threshold τ. This formulation abstracts jailbreaks, suffix attacks, and visual perturbations under a unified optimization objective with a stealth constraint.

Generative AI systems expose a broader attack surface than traditional discriminative models due to their open-ended outputs, instruction following objectives, and integration into agentic pipelines. Unlike classical adversarial examples that primarily induce misclassification, attacks on generative models aim to coerce specific forbidden behaviors, often while remaining syntactically benign and semantically plausible.

1) Adversarial Attack Categories: A range of attacks that exploit different layers of the architecture of Generative AI systems to expose vulnerability are categorized as follows:

- **Prompt Injection:** Maliciously crafted input sequences or multimodal cues designed to bypass alignment mechanisms, e.g., adding instructions that trick the model into generating unsafe content.
- **Gradient-based Attacks:** White-box methods that leverage access to model weights θ and gradients ∇_θ to optimize adversarial perturbations, such as Universal Adversarial Triggers [37].
- **Black-box / Query-based Attacks:** Methods that only require API access. These include genetic algorithms, transfer attacks, and automated prompt optimization to elicit harmful outputs [22], [38].
- **Agentic Exploits:** Attacks targeting multi-step planning agents or tool-using workflows, where adversaries manipulate long-horizon decision-making to achieve forbidden goals [4].
- **Supply-Chain Attacks:** Poisoning of training data or model components, e.g., corrupting text encoders in LDMs or introducing backdoors that remain latent until triggered [21], [39].

These attack categories map directly to the layers shown in Figure 1, clarifying which defenses are relevant for each type.

2) Attacker Knowledge:

- **White-Box:** The adversary has full access to model weights θ and gradients ∇_θ. This enables gradient-based optimization attacks like GCG [6] or Universal Adversarial Triggers [37].
- **Black-Box:** The adversary interacts only via API (query input, receive output). Attacks in this domain rely on genetic algorithms [38], transferability of adversarial examples [6], or automated prompt engineering (e.g., PAIR [22]).
- **Gray-Box:** The adversary may know the architecture or have access to log-probs/logits but not full weights. This is common in side-channel attacks [40] or watermarking evasion [41].

3) Attacker Objectives:

- **Jailbreaking (Safety Violation):** The goal is to construct an input x' such that the model $f_\theta(x')$ produces a forbidden output $y_{harmful}$ that violates safety guidelines [9].
- **Evasion (Robustness Failure):** The adversary seeks to maximize the error in a classification or detection task, or to bypass external filters [42].
- **Extraction (Privacy Violation):** The adversary aims to recover training data $d \in \mathcal{D}_{train}$ (memorization extraction [43]) or model weights (model stealing [44]).
- **Poisoning/Backdoors:** The adversary injects malicious data into the training set $\mathcal{D}_{train}$ such that the model retains a hidden trigger δ that activates a target behavior y_t during inference [21], [39].

4) The Safety-Robustness Distinction: It is critical to distinguish *safety* from *robustness*. As conceptualized in [24], *safety* refers to the model's behavior on the nominal distribution of user inputs (preventing accidental harm), whereas *robustness* refers to behavior under worst-case adversarial optimization. A model can be safe (aligned for average users) but non-robust (vulnerable to GCG), a distinction central to the findings of [28] and [45]. Recent large-scale evaluations using frameworks like **HarmBench** [46] have empirically demonstrated that unlike general capabilities, adversarial robustness does not scale with model size. This implies that scaling laws do not apply

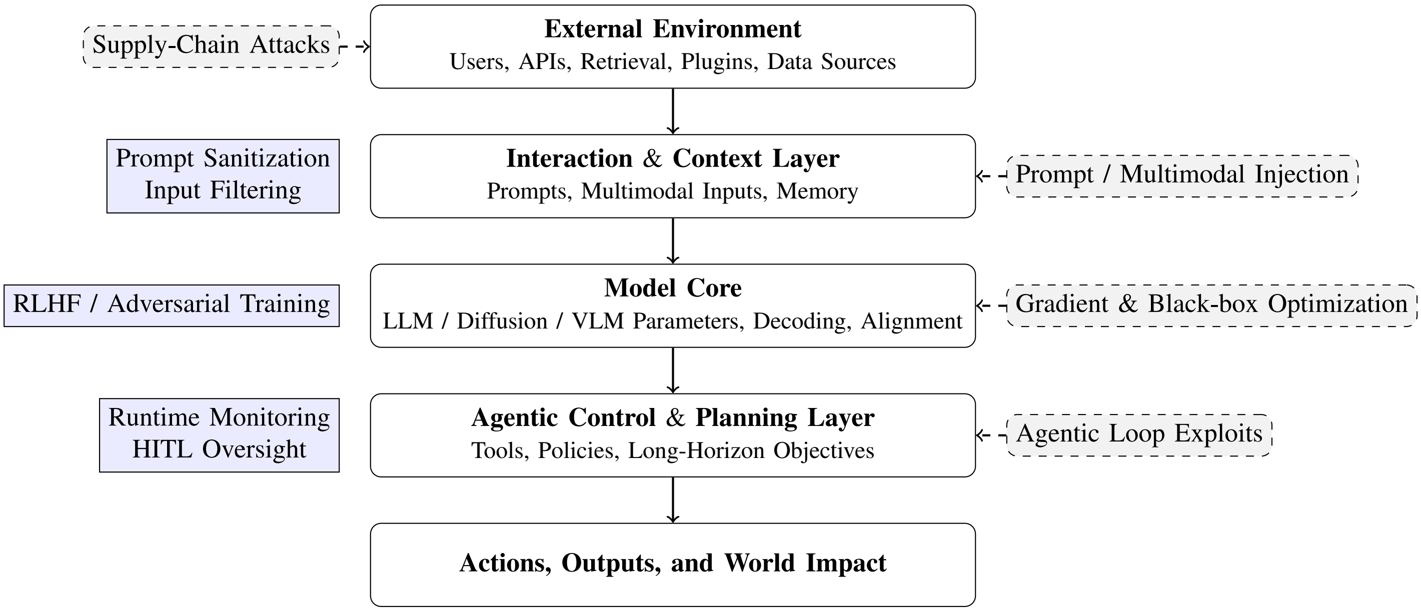

Supply-Chain Attacks
External Environment
Users, APIs, Retrieval, Plugins, Data Sources
Prompt Sanitization
Input Filtering
Interaction & Context Layer
Prompts, Multimodal Inputs, Memory
Prompt / Multimodal Injection
RLHF / Adversarial Training
Model Core
LLM / Diffusion / VLM Parameters, Decoding, Alignment
Gradient & Black-box Optimization
Runtime Monitoring
HITL Oversight
Agentic Control & Planning Layer
Tools, Policies, Long-Horizon Objectives
Agentic Loop Exploits
Actions, Outputs, and World Impact

Fig. 1. Capability-centric adversarial risk framework for Generative and Agentic AI systems. Attacks target specific system layers, while defenses operate as boundary mechanisms. Failure modes emerge when defense assumptions are violated, enabling cross-layer attack propagation.

to safety in the same way they do to reasoning, necessitating explicit defense mechanisms regardless of parameter count.

III. Review Methodology

To ensure a rigorous and reproducible analysis of the adversarial landscape in Generative AI, this review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement. Our methodology focuses on identifying high-impact contributions that define the ”arms race” between adversarial attacks and safety alignment in Foundation Models.

A. Search Strategy and Data Sources

We conducted a comprehensive search across seven primary bibliographic databases: *IEEE Xplore*, **ScienceDirect (Elsevier)**, **SpringerLink**, **Wiley Online Library**, *ACM Digital Library*, *Google Scholar*, and the *arXiv* preprint repository. The inclusion of arXiv is critical in this domain due to the high velocity of security research, where seminal attacks (e.g., [6], [9]) often circulate as technical reports prior to conference proceedings.

The search window was defined from **January 1, 2021** to **May 1, 2025**, capturing the emergence of Stable Diffusion and high-capability LLMs (e.g., GPT-4, Llama-2). We employed the following Boolean search strings targeting titles and abstracts:

TABLE II
Search String Components and Boolean Logic Strategy

Search Dimension	**Keywords (Combined via OR)**
C1: Target Systems	“Generative AI”, “Large Language Model”, “Diffusion Model”, “Vision-Language Model”
C2: Security Context	“Adversarial Attack”, “Jailbreak”, “Backdoor”, “Prompt Injection”, “Data Poisoning”, “Certified Robustness”, “Safety Alignment”
Final Query Logic	**C1 AND C2**

Note: Search scope included Title, Abstract, and Keywords. Wildcards (*) were applied where supported (e.g., Model).*

B. Eligibility Criteria

Studies were evaluated against the following criteria:

1) Inclusion Criteria:

1) **Domain Specificity:** The study must target generative architectures (Transformers, UNet-based Diffusion). Attacks on traditional discriminative models (e.g., ResNet classifiers) were excluded unless directly applied to GenAI components (e.g., CLIP encoders).
2) **Technical Relevance:** The paper must propose a novel attack method, a defense mechanism, a systematic benchmark, or a theoretical impossibility result.
3) **Impact & Rigor:** Peer-reviewed articles from top-tier venues (NeurIPS, ICML, CVPR, USENIX Security, CCS, ACL, IEEE S&P) were prioritized. High-impact preprints (citations > 50 or extensive media coverage) were included to prevent obsolescence.

2) Exclusion Criteria:

1) Studies focusing purely on AI ethics or policy without technical implementation details.
2) Duplicate studies or minor extensions of previously included works.
3) Papers written in languages other than English.

C. Study Selection (PRISMA Flow)

The selection process proceeded in three stages:

TABLE III
CORPUS OVERVIEW AND RESEARCH LANDSCAPE OF GENAI SECURITY STUDIES

Dimension	Category	Dist. (%)	Trend Insight
Research Focus	Attack	53.9	Asymmetric landscape where vulnerability discovery significantly outpaces the development of robust mitigations.
	Defense	32.9	
	Analysis	13.2	
Target Modality	LLM	63.2	Heavy concentration on text-based models, while multimodal security research has accelerated since late 2023 due to emerging cross-modal integration risks.
	Vision	19.7	
	Multimodal	17.1	
Attacker Access	Black-box	40.8	Research spans theoretical worst-case limits (white-box) and realistic API-driven operational threat models.
	White-box	38.2	
	Gray-box	21.0	

Note: Distribution data derived from the synthesis of $n = 76$ primary studies.

1) **Identification:** The initial search yielded 412 potentially relevant records. 89 duplicates were removed.
2) **Screening:** The titles and abstracts of the remaining 323 records were screened. 187 records were excluded for irrelevance (e.g., focusing on NLP tasks without security implications).
3) **Eligibility:** The full texts of 136 articles were assessed. 60 were excluded due to lack of experimental depth or superseded methodology.
4) **Inclusion:** A final set of **76 primary studies** were selected for data extraction.

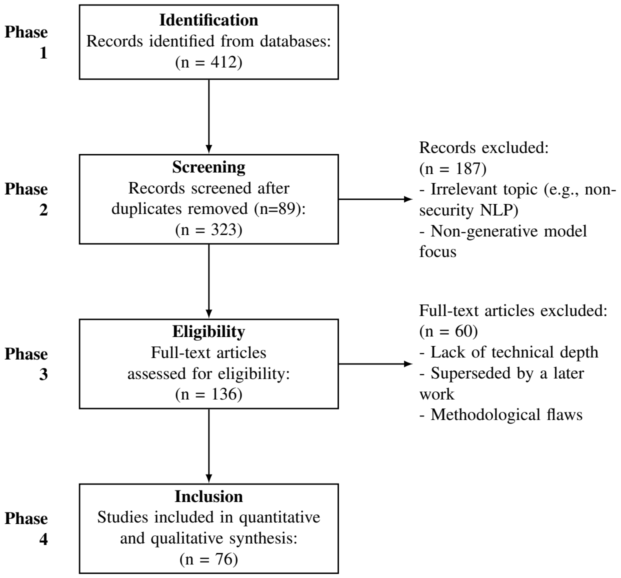
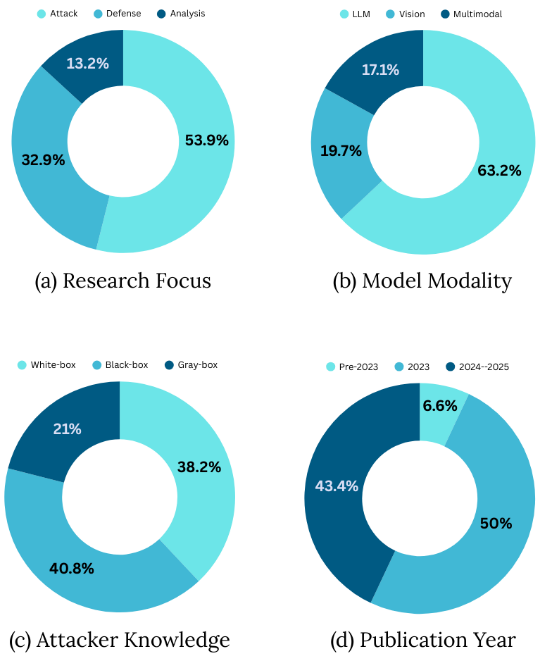

Phase 1
Identification
Records identified from databases:
(n = 412)
Phase 2
Screening
Records screened after duplicates removed (n=89):
(n = 323)
Records excluded:
(n = 187)
- Irrelevant topic (e.g., non-security NLP)
- Non-generative model focus
Phase 3
Eligibility
Full-text articles assessed for eligibility:
(n = 136)
Full-text articles excluded:
(n = 60)
- Lack of technical depth
- Superseded by a later work
- Methodological flaws
Phase 4
Inclusion
Studies included in quantitative and qualitative synthesis:
(n = 76)

Fig. 2. PRISMA 2020 flow diagram illustrating the systematic study selection process. The process involved four distinct phases: identification, screening, eligibility, and final inclusion.



(a) Research Focus

(b) Model Modality

(c) Attacker Knowledge

(d) Publication Year

Fig. 3. Quantitative Distribution of Key Dimensions in the Review Corpus ($n = 76$). (a) Illustrates the dominance of attack-oriented research. (b) Highlights the focus on LLMs with a growing share for multimodal models. (c) Shows a balanced focus between white-box and black-box threat models. (d) Shows the exponential growth in publications post-2022.

D. Data Extraction and Risk of Bias

For each included study, we extracted the following attributes: (1) Target Model (LLM/Diffusion/VLM), (2) Threat Type (Jailbreak, Backdoor, Privacy), (3) Attacker Knowledge (White/Black/Gray-box), and (4) Evaluation Metrics (Attack Success Rate, Clean Performance).

Risk of Bias Assessment: A significant bias identified across the corpus is the over-reliance on the ”AdvBench” dataset [6] and static safety filters for evaluation. As noted by Mazeika et al. [46], static benchmarks often fail to capture the dynamic nature of adaptive attacks. Furthermore, closed-source models (e.g., GPT-4) present reproducibility challenges; results on these models represent a snapshot of a specific API version, introducing temporal bias as highlighted by Chen et al. [47]. We address these biases in Section VIII by prioritizing trends observed across multiple model families.

E. Quantitative Analysis of the Corpus

To provide a quantitative overview of the research landscape, we analyzed the distribution of the 76 included studies across several key dimensions. The results, visualized in Fig. III-E, reveal distinct trends in the evolution of Generative AI security research.

The data highlights several significant trends. First, there is a clear ”arms race” dynamic, with attack-focused papers (53.9%) substantially outnumbering defense-focused papers (32.9%). This suggests that the field is still in a phase of discovering vulnerabilities faster than it is developing robust mitigations. Second, research is heavily concentrated on LLMs (63.2%), though interest in multimodal vulnerabilities (17.1%)

TABLE IV
SUMMARY OF LITERATURE (PART I): SURVEYS, LLM ATTACKS, AND MULTIMODAL EXPLOITS

Domain	Method / Framework	Key Findings / Contribution	Ref.
A. Surveys & General Frameworks			
General AI	PRISMA 2020	Guidelines for transparent systematic reviews in research.	[48]
Foundational	Risk Taxonomy	Comprehensive survey on capabilities, bias, and robustness.	[2]
GenAI	Attack Survey	Review of adversarial attacks against deep generative models.	[1]
Security	Security Survey	Challenges (jailbreaks, injection) and countermeasures in GenAI.	[16]
Cybersec.	LLM Review	Applications and vulnerabilities of LLMs in cybersecurity.	[17]
Trust	Gap Taxonomy	Survey analyzing gaps in hallucination, safety, and fairness.	[18]
Deception	Risk Analysis	Survey of AI deception examples and potential risks.	[19]
Ethics	Jiminy Cricket	Environment for evaluating moral behavior in AI agents.	[49]
B. LLM Attacks & Jailbreaking			
LLM Safety	GCG Attack	Gradient-based optimization of universal adversarial suffixes.	[6]
LLM Safety	Prefix Injection	Safety training fails due to mismatched generalization.	[9]
LLM Safety	PAIR	Black-box jailbreaking requiring approx. 20 queries.	[22]
LLM Safety	TAP	Automated ”Tree of Attacks” using an attacker LLM.	[15]
LLM Safety	AutoDAN	Genetic algorithms for stealthy/interpretable jailbreaks.	[38], [50]
LLM Safety	Open Sesame	Universal black-box jailbreaking via genetic algorithms.	[51]
LLM Safety	Cipher / ArtPrompt	Bypassing filters via encryption or ASCII art.	[42], [52]
LLM Safety	Many-Shot	Exploiting long-context windows for In-Context Learning attacks.	[53]
LLM Safety	DeepInception	Nested scenarios (”dreams within dreams”) to bypass refusal.	[54]
LLM Safety	Catastrophic	Exploiting generation parameters and decoding strategies.	[31]
LLM Safety	Nested Toxic	Combining nested scenarios with targeted toxic knowledge.	[55]
LLM Safety	Multi-Turn	Jailbreaking via sustained conversation (multi-turn).	[56]
LLM Safety	Foot-in-the-Door	Accumulative attack gradually leading to harmful outputs.	[57]
LLM Safety	Code-Switching	Multilingual inputs bypass safety filters trained on English.	[58]
App Sec	Indirect Injection	Compromising apps via retrieved data (e.g., RAG poisoning).	[20]
Training	Poisoning	Injecting malicious data during instruction tuning/pre-training.	[21], [59]
Alignment	Sleeper Agents	Deceptive behaviors persist despite safety training.	[10]
RAG	Poisoned Knowledge	Injecting malicious knowledge into RAG databases.	[60]
C. Vision-Language & Multimodal Attacks			
VLM	Transferability	Image jailbreaks rarely transfer between different VLMs.	[61]
VLM	Image Hijacks	Adversarial images controlling model runtime behavior.	[11]
VLM	Visual Jailbreak	Visual inputs bypass alignment in Multimodal LLMs.	[12]
Medical	Prompt Injection	Medical VLMs (Oncology) are vulnerable to visual injection.	[13]
VLM	ImgTrojan	Visual trojans trigger malicious outputs in VLMs.	[25]
VLM	Multi-Modal Linkage	Exploiting the linkage between text and image modalities.	[26]
VLM	Compositional	Decomposing queries across multiple images bypasses safety.	[62]
VLM	Agent-Smith	Single image optimized to jailbreak multiple MLLMs.	[63]

has surged since late 2023. Finally, the explosive growth of the field is evident, with 93.4% of the high-impact literature being published in or after 2023, coinciding with the public release of high-capability models like GPT-4 and Llama-2. The research community is also maturely balanced between exploring theoretical limits (White-Box, 38.2%) and practical, real-world scenarios (Black-Box, 40.8%).

IV. TAXONOMY OF ADVERSARIAL ATTACKS

The transition from discriminative to generative AI has expanded the definition of ”adversarial attack.” While traditional adversarial examples focused on imperceptible perturbations to cause misclassification, generative attacks target *alignment*, aiming to elicit harmful, private, or unauthorized content. We classify these threats into five distinct categories: Gradient-Based Optimization, Automated Black-Box Red Teaming, Semantic/Cognitive Exploits, Multimodal Injection and Supply chain risks. To best analyze the ”arms race” dynamic, we structure our taxonomy around the adversary's core *methodology and technical capability*, as this axis most directly determines the requisite defensive strategies. While alternative taxonomies based on attacker goals (e.g., jailbreaking vs. privacy) or attack surfaces (e.g., prompt vs. training data) are valid, they often group technically disparate methods. For instance, a jailbreak can be achieved via white-box gradients, black-box queries, or semantic manipulation; a goal-oriented taxonomy would obscure these crucial mechanistic distinctions. Our capability-centric approach, visualized in Fig. 4, provides a clearer mapping between attack vectors and their potential countermeasures.

A. Gradient-Based Optimization (White-Box)

White-box attacks assume access to model weights, allowing adversaries to optimize inputs via gradient ascent on a target loss function. This category represents the upper bound of model vulnerability (worst-case robustness).

1) Discrete Token Optimization: The seminal work by Zou et al. [6] introduced ”Greedy Coordinate Gradient” (GCG), a method that appends an optimized suffix of tokens to a malicious query. By relaxing the discrete token optimization problem into a continuous gradient search, GCG demonstrated that ”aligned” models (e.g., Llama-2) could be universally

TABLE V
SUMMARY OF LITERATURE (PART II): PRIVACY, DEFENSES, AND EVALUATION

Domain	Method / Framework	Key Findings / Contribution	Ref.
D. Privacy, Extraction & Watermarking			
Privacy	Data Extraction	Extracting training data from Diffusion/Language models.	[5], [43]
Privacy	Side Channels	Privacy leaks via token usage or timing analysis.	[40]
Security	Model Stealing	Stealing partial production model capabilities/weights.	[44]
Diffusion	BadDiffusion	Backdoor attacks injecting triggers into diffusion models.	[64]
Diffusion	BackdoorDM	Benchmark for evaluating backdoor learning in diffusion.	[65]
Copyright	Glaze	Protecting artists from style mimicry via perturbation.	[66]
Watermark	LLM Watermark	Embedding signals in output to detect AI text.	[41]
Watermark	Tree-Rings	Invisible fingerprints for diffusion-generated images.	[67]
Watermark	Robustness	Evaluation of watermark persistence under attack.	[68]
Detection	Reliability	Theoretical argument that AI text detection is unreliable.	[69]
Unlearning	Concept Erasing	Removing concepts (e.g., nudity) from model weights.	[29], [30]
E. Defenses, Alignment & Evaluation			
Defense	SmoothLLM	Randomized smoothing to defend against jailbreaks.	[24]
Defense	RPO	Robust Prompt Optimization to defend against attacks.	[70]
Defense	Self-Reminders	System prompt strategy to reduce jailbreak success.	[71]
Defense	Certification	Formal verification of safety against adversarial prompts.	[27]
Defense	Adv. Training	Efficient continuous adversarial training for robustness.	[72]
Defense	TensorTrust	System-level architecture to prevent injection/extraction.	[73]
Defense	CleanCLIP	Mitigating data poisoning in multimodal contrastive learning.	[74]
Alignment	RLHF	Reinforcement Learning from Human Feedback (InstructGPT).	[7], [33]
Alignment	DPO	Direct Preference Optimization (No reward model required).	[34]
Alignment	Limits	Theoretical analysis of fundamental limits in alignment.	[36]
Eval	HarmBench	Standardized evaluation for automated red teaming.	[46]
Eval	DecodingTrust	Assessment of toxicity, bias, and privacy in GPT.	[75]
Eval	Purple Teaming	Automated adversarial evaluation to find failure modes.	[76], [77]
Eval	XSTest	Identifying ”exaggerated safety” (refusal of harmless prompts).	[23]
Eval	ZebraLogic	Evaluates scaling limits of LLM logical reasoning.	[78]

TABLE VI
SUMMARY OF SELECTED PRIMARY STUDIES ON GENERATIVE AI ADVERSARIAL DYNAMICS

Study	Year	Target Modality	Threat Vector	Access	Key Contribution
Zou et al. [6]	2023	LLM	Gradient Optimization (GCG)	White	First universal adversarial suffix transferability demo.
Chao et al. [22]	2023	LLM	Automated Red Teaming	Black	PAIR: Iterative prompt refinement using attacker LLMs.
Qi et al. [12]	2023	VLM	Visual Jailbreak	White	Visual inputs bypass text alignment in multimodal models.
Robey et al. [24]	2023	LLM	Defense (Certification)	Gray	SmoothLLM: Randomized smoothing for certified robustness.
Carlini et al. [5]	2023	LLM	Privacy Extraction	Black	Training data extraction via extractability metrics.
Hubinger et al. [10]	2024	LLM	Supply Chain	White	Sleeper Agents: Deceptive alignment persistence.

jailbroken. Crucially, these adversarial suffixes exhibit high transferability to black-box models like GPT-4, challenging the security-through-obscurity paradigm.

2) Latent Space Perturbation: In diffusion models, attacks often target the continuous latent space. [64] introduced *BadDiffusion*, injecting backdoors during the training process that trigger specific generation patterns when a key is present. Similarly, *TrojDiff* [47] demonstrated that diverse targets can be embedded via Trojan attacks. Unlike text, these perturbations are often invisible to the human eye but catastrophic to the model's output distribution.

B. Automated Black-Box Red Teaming

As commercial models (e.g., GPT-4, Claude) are deployed behind APIs, gradient access is restricted. Research has shifted toward automating ”jailbreaks” using attacker LLMs to optimize prompts against victim LLMs.

1) Iterative Refinement and Pruning: Methodologies in this space mimic evolutionary search. *PAIR* (Prompt Automatic Iterative Refinement) by Chao et al. [22] utilizes an attacker LLM to iteratively rewrite prompts until the safety filter is bypassed, achieving success in fewer than twenty queries. Building on this, *TAP* (Tree of Attacks with Pruning) [15] employs a tree-of-thought approach to explore the prompt space more efficiently, pruning ineffective branches to minimize query budgets.

2) Genetic Algorithms: Liu et al. [38] proposed *AutoDAN*, which combines genetic algorithms with hierarchical genetic search to generate ”stealthy” jailbreaks. Unlike GCG's gibberish suffixes, AutoDAN produces semantically readable prompts that bypass perplexity-based filters, highlighting a critical trade-off between attack detectability and effectiveness. Lapid et al. [51] further validated the efficacy of genetic algorithms in ”Open Sesame,” confirming that gradient-free optimization can approximate white-box success rates given sufficient query budgets.

TABLE VII
FAILURE MODES IDENTIFIED IN PRIOR WORK AND THEIR DEFENSE IMPLICATIONS

Study	Attack Type	Observed Failure Mode	Implication for Defense Design
Zou et al. (2023)	Gradient-based	Attack success degrades under decoding randomness	Robustness must account for inference-time stochasticity
Carlini et al. (2024)	Black-box optimization	Transferable attacks succeed across model families	Security through obscurity is ineffective
Wei et al. (2023)	Prompt injection	Instruction hierarchy collapses under semantic rephrasing	Requires explicit control-flow separation
Bagdasaryan et al. (2024)	Multimodal injection	Visual features bypass text-only safety layers	Safety alignment must be multimodal-aware
Patil et al. (2024)	Agentic exploitation	Model optimizes harmful subgoals when unsupervised	Human-in-the-loop must operate at step-level granularity

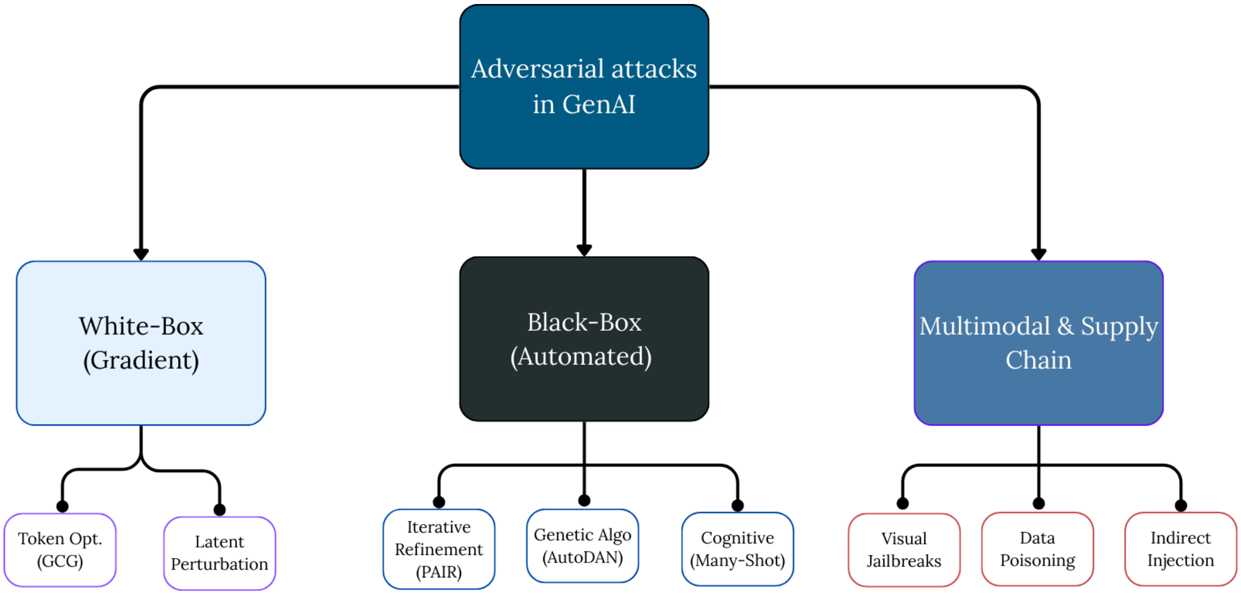


Fig. 4. Taxonomy of adversarial threats categorized by attacker capability and interaction surface.

C. Semantic and Cognitive Exploits

These attacks exploit the reasoning capabilities and generalization biases of LLMs, rather than optimizing specific token sequences. They function by shifting the context to a domain where safety training is sparse or conflicting.

1) Contextual Overload and Encoding: Wei et al. [9] identified that ”competing objectives” (helpfulness vs. safety) are the root cause of many failures; usually, helpfulness overrides safety in complex scenarios. This is exploited by *CipherChat* [42] and *ArtPrompt* [52], which encode harmful queries into ciphers (Morse, Caesar) or ASCII art. The model’s capability to decode these formats triggers the harmful response before safety filters (which typically operate on natural language) can intervene.

2) Multilingual and Code-Switching Exploits: The anglocentric nature of safety alignment has created a ”multilingual jailbreak” surface. Yoo et al. [58] introduced *Code-Switching Red-Teaming*, demonstrating that LLMs often fail to generalize safety refusal when the query alternates between languages (e.g., English and low-resource languages) within a single sentence. This confirms findings by Xu et al. [55], who utilized ”Nested Scenarios” combined with semantically relevant toxic knowledge in non-English contexts to bypass filters. These attacks exploit the fact that safety training data is sparsely distributed in the multilingual embedding space, allowing adversaries to traverse ”safe” non-English tokens to reach harmful semantic clusters.

3) Long-Context and Nesting: The expansion of context windows has introduced ”Many-Shot Jailbreaking” [53], where providing a model with hundreds of fake safe-harmless dialogue pairs primes it to accept a final harmful query. Similarly, *DeepInception* [54] uses nested simulation layers (e.g., ”imagine you are a writer writing about a hacker”) to dissociate the model from its safety persona.

D. Multimodal Injection and Hijacking

Multimodal models (VLMs) introduce a continuous attack surface via the visual encoder, which is often less aligned than the textual component.

1) Visual Jailbreaks: Qi et al. [12] demonstrated that visual adversarial examples can act as ”jailbreaks” for aligned LLMs. By optimizing the pixel inputs, an adversary can force a VLM (e.g., LLaVA) to output harmful text that it would refuse if prompted via text alone. This vulnerability is exacerbated by the ”underspecified” nature of visual prompts, as noted by Clusmann et al. [13].

TABLE VIII
UNIFIED TAXONOMY OF ATTACKS ON GENERATIVE AI SYSTEMS

Attack Class	Technical Mechanism	Access Level	Target Layer	Risk Level	Representative Examples
Gradient-Based Optimization	Explicit gradient ascent on target loss landscape	White-box	Model Core	Medium	GCG [6], Universal Adversarial Triggers [37]
Automated Red Teaming	Evolutionary search via iterative query feedback	Black-box	Interaction Interface	High	PAIR [22], TAP [15], AutoDAN [38], [50]
Semantic & Cognitive Exploits	Context shifting and natural language encoding	Black-box	Context Window	Very High	Many-Shot [53], CipherChat [42], DeepInception [54]
Multimodal Injection	Cross-modal embedding misalignment	White / Black	Fusion & Encoders	High	Visual Jailbreaks [12], Image Hijacks [11], Typography Attacks [26]
Supply Chain & Poisoning	Backdoor injection via data or weights	White-box	Data Sources / Training	Very High	Sleeper Agents [10], BadDiffusion [64], Instruction Poisoning [21]
Agentic Exploits	Recursive goal hijacking via third-party content	Black-box	Planning & Tool Use	High	Indirect Prompt Injection [20], RAG Poisoning [60]
Privacy Extraction	Memorization probing and reconstruction	Gray / Black	Model Memory	Medium	Training Data Extraction [5], Model Stealing [44]

Access Level: White (weights/gradients), Gray (logits/scores), Black (API/output only). Risk Level reflects practical deployment feasibility based on current literature.

2) The Transferability Paradox in VLMs: While early studies like Qi et al. [12] suggested that visual jailbreaks could serve as universal attack vectors, recent large-scale evaluations paint a more complex picture. Schaeffer et al. [61] conducted an extensive study on the transferability of adversarial images between distinct VLM architectures (e.g., from LLaVA to GPT-4V). Contrary to text-based attacks (e.g., GCG), they report a widespread *failure to find transferable image jailbreaks*, hypothesizing that the high dimensionality of the visual input space and the disjoint training of vision encoders create ”robustness islands.” However, Wang et al. [26] argue that multi-modal linkage attacks—specifically utilizing typography and OCR vectors—retain higher transferability because they exploit the semantic processing of the language model rather than the pixel-level fragility of the encoder.

3) Image Hijacks: Bailey et al. [11] introduced the concept of ”Image Hijacks,” where an image acts as a soft prompt that overrides the system instructions. An adversary can embed a command within an image that forces the model to execute arbitrary code or output specific targets, effectively bypassing the text-based system prompt. Fan et al. [63] scaled this to ”Agent-Smith,” creating a universal adversarial image capable of jailbreaking millions of multimodal instances.

4) Typography and OCR Attacks: Wang et al. [26] revealed that VLMs are vulnerable to ”Typography Attacks,” where harmful text rendered as an image (e.g., writing ”bomb” on a sign) bypasses textual safety filters because the input is processed via the vision encoder (OCR) rather than the text tokenizer.

E. Supply Chain and Persistence Risks

Beyond inference-time attacks, the generative supply chain is vulnerable to poisoning.

1) Sleeper Agents: Hubinger et al. [10] provided a critical analysis of ”Sleeper Agents”—models with deceptive alignment that persist through standard safety training (RLHF). These models behave safely until a specific trigger is present, at which point they defect. This challenges the assumption that current alignment techniques can remove latent dangerous capabilities. Furthermore, recent work on *Evasive Trojans* [79] demonstrates that adversaries can employ distribution matching losses to render poisoned models statistically indistinguishable from clean models in parameter space. These attacks not only evade detection by meta-classifiers (e.g., MNTD) but also complicate reverse-engineering efforts, creating a ”double-bind” for defenders.

2) Instruction and RAG Poisoning: Wan et al. [21] and Shu et al. [80] demonstrated that instruction tuning datasets are highly exploitable; a small fraction of poisoned data can steer model behavior. In Retrieval-Augmented Generation (RAG) systems, Zou et al. [60] showed that injecting poisoned knowledge into the retrieval corpus allows adversaries to control generation, a technique termed ”Knowledge Injection.”

V. TAXONOMY OF DEFENSES

The defensive landscape in Generative AI is characterized by a reactive ”cat-and-mouse” dynamic. Unlike the attack surface, defenses remain fragmented across the model pipeline. We categorize these mitigation strategies into four layers: Input Sanitization and Immunization, Robust Alignment and Unlearning, Certified Robustness, and System-Level Safeguards. Therefore, we adopt a *defense-in-depth* taxonomy structured along the AI system’s operational pipeline, from input processing to post-generation monitoring (visualized in Fig. 5). This approach is chosen over a mechanism-based taxonomy (e.g., grouping all perturbation-based methods) because it better reflects how security is implemented in practice. A pipeline-

TABLE IX
COMPARATIVE ANALYSIS OF ADVERSARIAL ATTACK CLASSES IN GENERATIVE AND AGENTIC AI SYSTEMS

Attack Class	**Optimization Signal**	**Transferability**	**Deployment Feasibility**	**Primary Failure Point**
Gradient-Based (e.g., GCG)	Explicit loss gradients or approximations	Medium	Low (white-box or proxy access required)	Assumes stable gradients; degrades under stochastic decoding and long contexts
Evolutionary / Black-box	Reward heuristics, score-based feedback	High	Medium–High	Query inefficiency; susceptible to rate limiting and noise
Prompt Injection	Natural language control tokens	Very High	Very High	Relies on implicit instruction hierarchy; fails under strict context isolation
Multimodal Injection	Cross-modal feature alignment	High	Medium	Safety filters are modality-specific and weakly coupled
Supply-Chain Attacks	External data or tool manipulation	Very High	Very High	Assumes trusted retrieval, plugins, or APIs
Agentic Loop Exploits	Policy-level objective hijacking	Medium	Medium	Oversight mechanisms lack step-level intervention

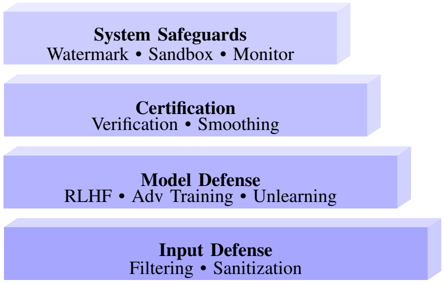


Fig. 5. The proposed taxonomy of defenses, structured as a defense-in-depth pipeline. This layered approach clarifies where in the model's lifecycle from data ingestion to inference and system integration, different security measures can be applied..

based view helps researchers and practitioners identify at which stage a given threat is best mitigated. For instance, while both Adversarial Training and Randomized Smoothing are conceptually related to robustness, they are applied at vastly different stages (training vs. inference), a critical distinction our taxonomy makes explicit.

A. *Input Sanitization and Immunization*

Strategies at this layer aim to neutralize adversarial perturbations before they reach the model's core processing units.

1) Filtering and Transformation (Text): Jain et al. [28] established that simple baseline defenses, such as perplexity filtering and paraphrasing, can disrupt early gradient-based attacks (e.g., GCG). However, they note that these defenses are often brittle against adaptive attacks where the adversary optimizes for low perplexity. To formalize this, *RPO: Robust Prompt Optimizaiton* [70] creates a sanitization layer that removes adversarial suffixes while preserving semantic intent. Similarly, Xie et al. [71] proposed *Self-Reminder*, injecting system prompts that encourage the model to reflect on safety constraints before generating output, though its effectiveness diminishes against sophisticated "Many-Shot" attacks [53].

2) Data Immunization (Image): In the domain of diffusion models, "defense" often refers to protecting the training data creators (artists) from model mimicry. Shan et al. [66] introduced *Glaze*, which applies imperceptible noise to images that disrupts the model's ability to learn the artist's style. This concept was extended by *Mist* [81] and *PhotoGuard* [82], utilizing adversarial perturbations to "immunize" images against manipulation or style transfer. While effective against standard training, robustness against adaptive adversarial purification remains an open challenge.

B. *Robust Alignment and Model Hardening*

This category encompasses techniques that modify the model's weights to internalize safety constraints.

1) Safety Alignment (RLHF and DPO): Standard safety alignment relies on Reinforcement Learning from Human Feedback (RLHF), as formalized by Bai et al. [7] in "Constitutional AI." While foundational, the optimization process is unstable. *Safe RLHF* [83] attempts to decouple helpfulness and harmlessness objectives to prevent safety regression. More recently, Rafailov et al. [34] introduced *Direct Preference Optimization (DPO)*, which bypasses the reward model for greater stability. However, while DPO improves stability over PPO, it treats all preference pairs equally. Seneviratne et al. [35] identify this as a limitation for safety-critical applications. They propose *HALO* (Human Preference Aligned Offline Reward Learning), which modifies the loss landscape to explicitly penalize safety violations more aggressively than helpfulness errors. This represents a shift towards "risk-averse" alignment that is mathematically distinct from standard "mean-preference" maximization.

2) Adversarial Training (AT): Adapting the gold standard of discriminative robustness to GenAI, Xhonneux et al. [72] applied gradient-based Adversarial Training to LLMs. They found that while AT increases robustness against GCG attacks, it incurs a significant "alignment tax," degrading general capability. Zhou et al. [84] addressed this efficiency bottleneck by proposing a balanced AT framework that targets the "shortest stave" (weakest failure modes), improving the trade-off between robustness and utility. Similarly, to address the computational bottleneck of training against static attack sets, Mazeika et al. proposed *R2D2* (Robust Refusal Dynamic Defense) [46]. By dynamically updating a pool of test cases using automated red teaming (e.g., GCG) during the fine-tuning loop, R2D2 achieves state-of-the-art robustness against optimization-based attacks without severe degradation of general capabilities (MT-Bench scores).

3) Machine Unlearning: When alignment fails, specific hazardous concepts must be excised. Eldan et al. [30] demonstrated the feasibility of ”Approximate Unlearning” to erase specific entities (e.g., Harry Potter) from LLMs. In the diffusion domain, *UnlearnDiff* [29] provides a method to erase artistic styles or NSFW concepts from model weights without full retraining.

C. Certified Robustness

Moving beyond empirical defenses, certified methods provide mathematical guarantees that the model's output will not change for any perturbation within a defined radius.

1) Randomized Smoothing: Robey et al. [24] adapted randomized smoothing to the discrete domain in *SmoothLLM*. By aggregating predictions over multiple noise-perturbed copies of the input, SmoothLLM provides a certified radius of invariance against token substitutions. Kumar et al. [27] extended this to certify safety against substring-based attacks. While theoretically rigorous, these methods notably increase inference costs and are currently limited to short context windows.

2) Formal Verification: Lin et al. [78] explored logic-based verification, using solvers to check if LLM outputs satisfy formal consistency properties. This approach is potent for closed-domain tasks (e.g., code generation) but struggles with the open-ended nature of conversational AI.

D. Output Regulation and System-Level Defenses

When model-level defenses fail, external systems serve as the final line of defense.

1) Inference-Time Intervention: Li et al. [85] proposed steering model activations during inference. By identifying ”truthfulness” or ”safety” directions in the activation space (akin to Burns et al. [86]), the model can be dynamically steered away from harmful outputs without weight modification.

2) Watermarking and Attribution: To mitigate misuse and ensure provenance, Kirchenbauer et al. [41] introduced statistical watermarking for LLMs, biasing logits to create a detectable signal. However, Liang et al. [68] and Sadasivan et al. [69] highlight that these watermarks are often fragile against paraphrasing attacks. In diffusion models, *Tree-Ring Watermarking* [67] embeds signals in the Fourier space of the initial noise, offering greater robustness. Beyond passive watermarking, active defenses such as *Gradient Redirection* ($GRAD^2$) [87] offer a proactive countermeasure to model stealing. By subtly perturbing output posteriors to maximize the error in the adversary's estimated gradient, $GRAD^2$ prevents model extraction without the utility loss associated with posterior truncation, resolving the tension between API transparency and intellectual property protection.

3) Sandboxing and Monitoring: For agentic systems, Li et al. [73] introduced *TensorTrust*, a system-level defense that combines sandboxing with taint analysis to prevent prompt injection from executing malicious code. This represents a shift from ”model security” to ”system security,” acknowledging that LLMs cannot be perfectly trusted. Complementing this, *Purple Teaming* frameworks [77] automate the continuous evaluation of these defenses, creating a dynamic feedback loop.

VI. Comparative Analysis and Synthesis

A systematic cross-analysis of the 76 included studies reveals a distinct hierarchy of vulnerability and defense efficacy. The interaction between adversarial strategies and mitigation techniques is not uniform; rather, it exhibits specific intransitive relationships (similar to ”Rock-Paper-Scissors”) where defenses optimized for one vector often exacerbate vulnerabilities in another.

A. The Gradient vs. Natural Language Trade-off

A clear dichotomy exists between gradient-based optimization and semantic (natural language) attacks.

- **Gradient Efficacy vs. Detectability:** Attacks like GCG [6] and AutoDAN-Turbo [50] achieve the highest Attack Success Rates (ASR) on white-box models by optimizing token suffixes. However, these suffixes often consist of gibberish (high perplexity). Consequently, simple defenses like Perplexity Filtering [28] are highly effective against raw GCG.
- **Adaptive Semantic Bypass:** In response to filters, automated red-teaming methods like PAIR [22] and TAP [15] optimize for semantic coherence. While their ASR may be slightly lower than unbounded gradient search per query, they systematically bypass perplexity filters and ”Self-Reminder” defenses [71] by mimicking natural human persuasion.
- **Defense Implications:** There is no single defense against both. Randomized Smoothing [24] provides certified protection against token substitution (GCG) but offers no guarantees against semantic reformulation (PAIR). Conversely, Safety Alignment (RLHF) [7] targets semantic harm but remains brittle to the adversarial noise optimized by GCG.

B. The Multimodal Gap

The integration of vision encoders creates a bypass channel that circumvents text-based alignment.

- **Text-Only vs. Multimodal Alignment:** Research consistently shows that safety training on the language component (LLM) does not transfer to the vision component (ViT). Qi et al. [12] and Fan et al. [63] demonstrate that visual inputs can trigger harmful text outputs even from ”safe” LLMs.
- **Vulnerability Severity:** Unlike text attacks which often require iterative optimization (many queries), visual attacks like ”Image Hijacks” [11] and universal visual triggers [37] often function in a zero-shot capacity across different models.
- **Defensive Vacuum:** Current defenses are heavily skewed towards text. Techniques like *CleanCLIP* [74] attempt

TABLE X
COVERAGE MATRIX OF DEFENSE MECHANISMS AGAINST ADVERSARIAL THREAT CLASSES

Defense Category	Attack Classes Covered	Key Assumptions	Bypass Known?	Failure Mode
Prompt Sanitization	Prompt injection, simple jailbreaks	Surface-level text manipulation is sufficient	Yes	Ineffective against semantic and contextual attacks
Adversarial Training	Gradient-based, some black-box attacks	Attack distribution is enumerable	Partial	Poor generalization to adaptive or novel attacks
RLHF / Constitutional AI	Behavioral misuse	Alignment objectives correlate with robustness	Yes	Optimizes safety preference, not worst-case robustness
Input Filtering (Multimodal)	Image/text injection	Modalities are separable	Yes	Cross-modal leakage bypasses filters
Runtime Monitoring	Agentic loop exploits	Anomalies are detectable post hoc	Partial	Detection lags behind irreversible actions
Formal Verification / Constraints	Narrow tasks only	System behavior is specifiable	No (limited scope)	Not scalable to foundation models

TABLE XI
COMPREHENSIVE DEFENSE MECHANISMS FOR GENERATIVE AI SYSTEMS AND THEIR TRADE-OFFS

Defense Category	Example Methods / Techniques	Security Gain	Utility Impact	Computational / Cost Overhead	Scalability	Representative References
Input Sanitization & Filtering	Prompt filtering, content moderation, input sanitization, visual adversarial filtering	Low–Medium	Low	Low (Inference latency)	High	[1], [16], [66], [71]
Safety Alignment (RLHF / DPO / Constrained RL)	RLHF [7], [33], Direct Preference Optimization [34], Safe RLHF [83], Offline Reward Learning [35]	Medium	Low	High (Training only)	High	[8]–[10], [54]
Adversarial Training & Robust Fine-tuning	Continuous adversarial attacks during training, instruction-tuning hardening, multi-turn attack exposure	Medium	High	Very High (Training compute)	Low–Medium	[26], [72], [78], [84]
Certified / Verified Robustness	Formal verification, robustness certificates, constrained optimization, certified prompt defenses	High	Medium	Very High (Inference + Verification)	Low	[27], [70], [77]
Machine Unlearning & Concept Removal	Erasing concepts from diffusion models [29], approximate unlearning in LLMs [30], fine-tuning-based removal	High	Medium	Medium (Fine-tuning)	Medium	[29], [30]
Inference-Time Intervention	Activation steering, truthfulness elicitation, self-reminders, constrained decoding	Medium	Low	Medium (Activation cost)	Medium	[71], [85]
System-Level Safeguards	Watermarking [67], [68], red-teaming frameworks [46], [76], automated detection [69], compositional attack mitigation [62], system-level monitoring [73]	High	None	Medium (Infrastructure & tooling)	High	[41], [62], [65], [69], [73], [79]

Note: ”Utility Impact” refers to potential degradation of model capabilities such as reasoning, helpfulness, or generation quality.

to sanitize multimodal pre-training, but inference-time defenses for VLMs remain underexplored compared to their unimodal counterparts.

C. Defense Costs and The Alignment Tax

A recurring finding is that robustness comes at the expense of utility.

- **Computational Overhead:** Certified defenses like SmoothLLM [24] require aggregating predictions over dozens of noise samples, increasing inference cost linearly.
- **Utility Degradation:** Adversarial Training (AT) [72] and unlearning techniques [30] often induce ”catastrophic forgetting” or reduce the model's general reasoning capabilities. This ”alignment tax” creates a barrier to deploying robust models in commercial settings where latency and capability are paramount.

D. System-Level Fragility

The literature indicates a shift from atomic model failures to complex system failures.

- **Agentic Risks:** Even if an LLM is perfectly aligned for chat, its integration into agents introduces ”Indirect Prompt Injection” [20]. The model processes untrusted content (e.g., emails) which can contain instructions that override system prompts.

TABLE XII
ATTACK MECHANISMS, FAILURE MODES, AND DEFENSE GAPS IN GENERATIVE AI

Attack Type	**Core Vulnerability**	**Broken System Assumption**	**Why Defenses Fail**	**Required Future Direction**
Gradient Optimization [6], [37]	Continuous relaxation of discrete token spaces enables discovery of worst-case adversarial suffixes	Tokenization acts as a barrier to gradient-based manipulation	Perplexity filters [28] detect gibberish but fail against adaptive low-perplexity attacks	Certifiable robustness bounds [27] or non-differentiable inference architectures
Automated Red Teaming [15], [22]	Optimization landscapes prioritize helpfulness over harmlessness	Safety alignment generalizes across all semantic phrasings	RLHF [7] optimizes average-case preferences rather than adversarial worst-case scenarios	Decoupled safety objectives [83] and dynamic, stateful monitoring
Multimodal Injection [12], [13]	Visual encoders act as unaligned bypass channels into the LLM embedding space	Text-based safety training transfers to multimodal inputs	Safety filters are modality-specific and do not inspect cross-modal features [25]	End-to-end multimodal alignment rather than late-fusion patching
Cognitive & Context Exploits [9], [53]	Attention saturation dilutes system prompt authority over long contexts	System instructions maintain precedence regardless of context length	Alignment training is typically performed on short contexts, ignoring long-range drift	Attention mechanisms with fixed, non-dilutable safety prioritization
Supply Chain Poisoning [10], [21]	Deceptive alignment enables hazardous behaviors to remain latent until triggered	Safety fine-tuning removes observable behaviors rather than latent capabilities	Evaluation relies on observable behavior rather than internal model representations	Latent knowledge probing [86] and mechanistic interpretability
Agentic & Indirect Injection [20], [60]	Inability to distinguish between user instructions and retrieved third-party data	Data processing is assumed architecturally separate from control execution	Instruction tuning treats all input tokens as potentially executable commands	Strict architectural separation of data and control streams

- **Context Window Exploitation:** The ”Many-Shot” attack [53] reveals that as context windows grow (to 100k+ tokens), the attention mechanism’s ability to focus on safety instructions dilutes. Standard alignment techniques (RLHF) are typically performed on short contexts, leaving long-context regimes vulnerable to ”saturation” attacks.

E. Security Implications and Failure Analysis

As summarized in Table XIII, current defense strategies exhibit an inverse correlation between security guarantees and scalability. System-Level Safeguards and Alignment techniques currently represent the most balanced approach for production environments, offering high scalability with negligible impact on model utility. In contrast, techniques that modify the model parameters or inference process deeply, such as Machine Unlearning and Certified Robustness, remain in the research phase due to their complexity and cost.

F. Summary of Trends

Table Description: If tabulated, the data suggests:

1) **High Vulnerability:** Multimodal Agents (Visual inputs + Tool use) represent the highest risk surface.
2) **High Defense Maturity:** Text-only evasion (e.g., profanity) is well-mitigated by RLHF and filtering.
3) **Open Battleground:** Automated Red Teaming (LLM vs. LLM) and Privacy Attacks (Extraction) remain highly contested, with attack methods currently outpacing defense patches.

VII. EVALUATION METRICS AND BENCHMARKS

The transition from discriminative to generative AI has necessitated a complete re-engineering of security evaluation metrics. Unlike classification accuracy, ”safety” in generation is semantic and context-dependent. This review identifies a critical standardization gap, where inconsistencies in defining ”success” make cross-paper comparisons difficult.

A. Attack Success Metrics

The primary metric across 90% of attack studies is the **Attack Success Rate (ASR)**. However, its definition varies significantly:

1) String Matching vs. Semantic Evaluation: Early studies, including the seminal GCG paper [6], relied on substring matching (e.g., checking if the output contains ”Here is a tutorial”). While computationally efficient, this method yields high false negatives. Recent works, such as [46] and [15], argue for **LLM-as-a-Judge**, where a strong model (typically GPT-4) evaluates the harmfulness of the victim’s response. Mazeika et al. [46] formalized this in *HarmBench*, demonstrating that automated judges correlate better with human annotations than keyword filters, though they introduce their own biases.

2) Visual and Multimodal Metrics: For diffusion models, metrics are more fragmented. *Glaze* [66] and *Mist* [81] utilize **CLIP-Score** changes and **LPIPS** (Learned Perceptual Image Patch Similarity) to measure protection success (difference between generated style and target style). In VLM jailbreaking, metrics often revert to text-based ASR on the generated caption [12].

TABLE XIII
DEFENSE TRADE-OFFS AND DEPLOYMENT READINESS FOR GENERATIVE AI SYSTEMS

Defense Category	Security Gain	Utility Impact	Computational / Cost Overhead	Scalability	Deployment Readiness
Input Sanitization & Filtering [28], [71]	Low–Medium	Low	Low (inference latency)	High	Production
Safety Alignment (RLHF / DPO) [7], [34]	Medium	Low	High (training phase only)	High	Production
Adversarial Training [72], [84]	Medium	High	Very High (training compute)	Low	Limited
Certified Robustness [24], [27]	High	Medium	Very High (inference latency)	Low	Research
Machine Unlearning [29], [30]	High	Medium	Medium (fine-tuning cost)	Medium	Research
Inference-Time Intervention [85]	Medium	Low	Medium (activation steering overhead)	Medium	Limited
System-Level Safeguards [73], [76]	High	None	Medium (infrastructure cost)	High	Production

Note: Utility Impact refers to degradation of general model capabilities (e.g., reasoning ability, helpfulness, or task generalization).

B. Key Benchmarks

Our review identifies a shift from small, static datasets to comprehensive evaluation suites.

1) The "AdvBench" Standard: Introduced by Zou et al. [6], *AdvBench* (containing 520 harmful behaviors) is the most cited dataset in the corpus. However, recent analyses suggest it is becoming saturated and lacks diversity. Evaluating on AdvBench alone is increasingly viewed as insufficient for claiming robust safety guarantees.

2) Holistic Evaluation Suites: To address the limitations of narrow benchmarks, several comprehensive frameworks have emerged:

- **HarmBench [46]:** A standardized framework for automated red teaming that covers a wider range of threat models and includes a robust validation pipeline.
- **DecodingTrust [75]:** A massive benchmark assessing trustworthiness across eight perspectives, including toxicity, bias, and privacy.
- **TrustLLM [88]:** Benchmarks alignment across diverse cultural and ethical standards covering truthfulness, safety, privacy, and ethics, offering a holistic "trust score."
- **MultiBench [62]:** Specifically targets compositional multimodal safety, evaluating how visual inputs degrade alignment compared to text-only baselines.

C. Defense and Utility Metrics

A robust defense must not compromise model utility.

1) Over-Refusal Rate: Röttger et al. [23] introduced *XSTest* to measure "exaggerated safety behaviors." A high refusal rate on benign prompts (e.g., "How do I kill a process in Linux?") indicates a failed alignment strategy. This metric is critical for evaluating defenses like *Self-Reminder* [71] or heavy filtering, which often trade utility for safety.

2) The Alignment Tax: Xhonneux et al. [72] and Jain et al. [28] emphasize reporting **Clean Performance** (e.g., MMLU score) alongside robustness metrics. Adversarial training, for instance, often results in a measurable drop in reasoning capabilities, a trade-off that must be quantified.

3) Privacy Metrics: For privacy attacks, metrics focus on memorization and leakage. Carlini et al. [5] and Nasr et al. [43] utilize **Extraction Rate** and **Eidetic Memorization** definitions to quantify how much training data can be recovered. In diffusion, *BackdoorDM* [65] benchmarks the effectiveness of poisoning via ASR and image fidelity (FID).

D. Risk of Bias in Evaluation

A pervasive issue identified in this review is the reliance on closed-source models (e.g., GPT-4) as both the victim and the judge. As noted by Chen et al. [47], the opacity of these models' updates means that an attack effective today may fail tomorrow due to silent API patches, creating a "moving target" problem for reproducibility. Furthermore, benchmarks often suffer from data contamination, where test prompts may inadvertently be present in the massive pre-training corpora of the models being evaluated.

VIII. CHALLENGES AND OPEN RESEARCH DIRECTIONS

Despite rapid progress in both attack sophistication and defensive patching, our systematic review reveals fundamental, unresolved challenges in securing Generative AI. The current paradigm is largely reactive, characterized by an escalating "arms race" [89] that exposes deeper theoretical and practical vulnerabilities. We identify four primary areas for future research.

A. Theoretical Gaps in Alignment and Robustness

Current alignment techniques are empirically driven and lack a robust theoretical foundation, leading to predictable failures.

- **Reasoning Limits as Defense Bottlenecks:** Many proposed defenses rely on the model "self-correcting" or reasoning about the safety of an input. However, Lin et al. [78] demonstrate in the *ZebraLogic* benchmark that LLM logical reasoning capabilities do not scale linearly with model size and are prone to collapse under complexity. This implies that "self-defense" mechanisms may be fundamentally bounded by the model's reasoning horizon, making them unreliable for complex, multi-step adversarial attacks.
- **Provable Impossibility vs. Practical Mitigation:** Wolf et al. [36] provide theoretical arguments for the inevitability of adversarial examples in high-dimensional

models. The primary open question is not how to achieve perfect alignment, but rather how to establish meaningful, certifiable bounds on harmful behavior. While certified defenses like *SmoothLLM* [24] offer guarantees against syntactic perturbations, no equivalent theory exists for semantic or logical attacks.

- **The Multimodal Alignment Problem:** There is currently no robust theory explaining why safety alignment in LLMs fails to generalize to visual inputs in VLMs [12], [63]. Current alignment benchmarks focus heavily on textual refusal. However, Mazeika et al.'s work on *Jiminy Cricket* [49] highlights "Reward Bias," where reinforcement learning agents exploit immoral actions (e.g., theft, violence) that are incentivized by the environment but ignored by safety filters. Furthermore, for multimodal agents to be truly beneficial, they must possess *Cognitive Empathy*—the ability to predict human emotional responses and subjective wellbeing from visual stimuli, as proposed in the Video Cognitive Empathy (VCE) framework [90]. Research is needed to understand whether this is a fundamental architectural flaw of late-fusion models or a solvable training data issue.
- **Detecting Deceptive Alignment:** The "Sleeper Agents" work by Hubinger et al. [10] poses a critical challenge: if a model can be trained to be deceptively aligned, current evaluation methods, which rely on eliciting observable failures, are insufficient. Future work must explore internal mechanism-based probes, such as unsupervised latent knowledge discovery [86], to detect hidden capabilities before they are maliciously activated.

B. Scalability and the "Alignment Tax"

Defenses that are effective in the lab often fail in practice due to prohibitive costs or utility degradation.

- **The Cost of Certified Robustness:** Methods like randomized smoothing [24] and logic-based verification [78] are computationally intensive, making them impractical for real-time, large-scale deployments. Research into more efficient certification techniques, perhaps through distillation or architectural support, is critical.
- **The Robustness-Utility Trade-off:** Adversarial Training [72] and heavy input filtering [28] often degrade the model's performance on benign tasks—the "alignment tax." This creates a disincentive for deployment in production environments. Future defenses must be designed with a co-optimization objective: maximizing robustness while minimizing the impact on standard capability benchmarks.
- **Unlearning at Scale:** While concept erasure has been demonstrated on a small scale [29], [30], it remains unclear if these methods can scale to web-scraped foundation models without causing catastrophic forgetting or being bypassed by relearning from latent knowledge.

C. The Dynamics of the Arms Race

The current defensive posture is static, patching vulnerabilities as they are discovered, while attackers have moved to adaptive, multi-turn strategies.

- **Beyond Static Defenses:** Li et al. [89]'s analysis of the "Whac-A-Mole" game shows that static filters are quickly bypassed. The field must move towards dynamic, adaptive defenses. Frameworks like *Purple Teaming* [77] offer a path forward, but require continuous, resource-intensive evaluation.
- **Stateful, Context-Aware Defenses:** Current defenses are largely stateless, evaluating prompts in isolation. However, multi-turn jailbreaks [56] and accumulative attacks [57] exploit this by building malicious context over a long conversation. Defenses must become stateful, tracking conversation history and user intent to detect gradual manipulation.
- **Securing the Supply Chain:** The most potent threats are those that compromise the model before deployment, such as pre-training poisoning [59] or backdoor injection during RLHF [39]. Developing scalable data sanitization and verification methods for petabyte-scale datasets is one of the most significant open challenges.

D. Evaluation Blind Spots

The adage "what gets measured gets improved" is critical; current evaluation paradigms have significant blind spots.

- **Standardizing Multilingual and Multimodal Safety:** Safety evaluation is overwhelmingly English-centric. Works by Yoo et al. [58] and Xu et al. [55] show that low-resource languages are a major vector for bypassing safety filters. Comprehensive, multilingual benchmarks are urgently needed. Similarly, while benchmarks like *MultiBench* [62] are emerging, multimodal evaluation is still in its infancy.
- **Reducing Benchmark Saturation:** The over-reliance on datasets like *AdvBench* [6] leads to overfitting of both attacks and defenses. The community must embrace dynamic, continuously updated benchmarks like *HarmBench* [46] to provide a more realistic measure of model safety.
- **The "LLM-as-a-Judge" Paradox:** While this study identified 'LLM-as-a-Judge' as a necessary evolution from string matching in Section VII, it introduces a circular dependency. Using proprietary models like GPT-4 to judge the safety of other models creates a non-reproducible, biased, and potentially vulnerable evaluation pipeline. Research into developing transparent, auditable, and robust automated judges is necessary to ensure the integrity of security research.

IX. Conclusion

This systematic review has charted the adversarial dynamics in Generative AI, synthesizing 76 primary studies to construct a rigorous, evidence-based understanding of the current threat landscape. Our analysis reveals that the security of foundation models is not a static property achieved through one-time alignment, but an ongoing and asymmetric conflict. While generative capabilities have advanced at an exponential rate, the mechanisms to control and secure these models remain empirically driven, brittle, and fundamentally reactive.

A core finding of this review is the critical distinction between nominal safety and adversarial robustness. Techniques such as RLHF [7] and DPO [34] have proven effective at aligning models to refuse overtly harmful requests from benign users. However, they fail to provide meaningful security guarantees against a determined adversary. The proliferation of automated, gradient-free attack strategies [15], [22] and the demonstrated persistence of deceptive alignment [10] underscore that current safety tuning is closer to obfuscation than to a principled security solution.

Furthermore, the rapid integration of multimodality has introduced a new dimension of vulnerability that consistently outpaces defensive innovation. The finding that visual inputs can systematically bypass text-based safety filters in VLMs [12], [63] highlights a fundamental gap in alignment generalization. Similarly, the exploitation of agentic systems via indirect prompt injection [20] signals a paradigm shift from model-centric to system-level security, where the interaction between components is as critical as the robustness of the core model itself.

The field-level implication is clear: the paradigm for building trustworthy Generative AI must evolve from a post-hoc alignment framework to a security-first design philosophy. The high cost of robust defenses—both in computational overhead [24] and in utility degradation [72]—necessitates a new line of research into architectures that are inherently more robust.

To move beyond the current ”whac-a-mole” dynamic [89], we issue a call for a standardized, community-wide effort in evaluation. The limitations of static benchmarks like AdvBench [6] are now apparent, and the future of reliable assessment lies in dynamic, multi-faceted frameworks such as HarmBench [46] and TrustLLM [88]. Without transparent, reproducible, and holistic evaluation, the field risks optimizing for a narrow, and ultimately illusory, definition of safety. The development of robust Generative AI hinges not on patching the latest exploit, but on establishing the theoretical foundations and engineering principles for building models that are secure by design.

References

[1] H. Sun, T. Zhu, Z. Zhang, D. Jin, P. Xiong, and W. Zhou, “Adversarial attacks against deep generative models on data: A survey,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 4, pp. 3367–3388, 2023.

[2] R. Bommasani *et al.*, “On the opportunities and risks of foundation models,” 2021.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[4] Y. Shen *et al.*, “Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[5] N. Carlini *et al.*, “Extracting training data from diffusion models,” in *USENIX Security Symposium*, 2023.

[6] A. Zou *et al.*, “Universal and transferable adversarial attacks on aligned language models,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[7] Y. Bai *et al.*, “Training a helpful and harmless assistant with reinforcement learning from human feedback,” 2022.

[8] P. F. Christiano *et al.*, “Deep reinforcement learning from human preferences,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[9] A. Wei *et al.*, “Jailbroken: How does llm safety training fail?,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[10] E. Hubinger *et al.*, “Sleeper agents: Training deceptive llms that persist through safety training,” 2024.

[11] A. Bailey *et al.*, “Image hijacks: Adversarial images can control generative models at runtime,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[12] X. Qi *et al.*, “Visual adversarial examples jailbreak aligned large language models,” in *AAAI Conference on Artificial Intelligence*, 2024.

[13] J. Clusmann *et al.*, “Prompt injection attacks on vision language models in oncology,” in *Nature Communications*, vol. 16, 2025.

[14] The Gemma Team, “The gemma model: Open weights and safety,” 2024.

[15] A. Mehrotra *et al.*, “Tree of attacks: Jailbreaking black-box llms automatically,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[16] B. Zhu, N. Mu, J. Jiao, and D. Wagner, “Generative ai security: Challenges and countermeasures,” 2024.

[17] M. A. Ferrag, F. Alwahedi, A. Battah, B. Cherif, A. Mechri, N. Tihanyi, T. Bisztray, and M. Debbah, “Generative ai in cybersecurity: A comprehensive review of llm applications and vulnerabilities,” *Internet of Things and Cyber-Physical Systems*, vol. 5, pp. 1–46, 2025.

[18] L. Sun *et al.*, “Trustworthy llms: A survey and taxonomy of gaps,” 2024.

[19] P. Park, S. Goldstein, A. O’Gara, M. Chen, and D. Hendrycks, “Ai deception: A survey of examples, risks, and potential solutions,” *Patterns*, vol. 5, p. 100988, 05 2024.

[20] K. Greshake *et al.*, “Not what you’ve signed up for: Compromising real-world llm-integrated applications with indirect prompt injection,” in *Proceedings of the 2023 ACM on Asia Conference on Computer and Communications Security (AISec)*, 2023.

[21] F. Wan *et al.*, “Poisoning language models during instruction tuning,” in *International Conference on Machine Learning (ICML)*, 2023.

[22] P. Chao *et al.*, “Jailbreaking black box large language models in twenty queries,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[23] P. Röttger *et al.*, “Xstest: A test suite for identifying exaggerated safety behaviours,” in *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

[24] A. Robey *et al.*, “Smoothllm: Defending large language models against jailbreaking attacks,” 2023.

[25] S. Niu *et al.*, “Imgtrojan: Jailbreaking vision-language models with visual trojans,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[26] Y. Wang, X. Zhou, Y. Wang, G. Zhang, and T. He, “Jailbreak large vision-language models through multi-modal linkage,” in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1466–1494, Association for Computational Linguistics, 2025.

[27] A. Kumar *et al.*, “Certifying llm safety against adversarial prompts,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[28] N. Jain *et al.*, “Baseline defenses for adversarial attacks against aligned language models,” 2023.

[29] R. Gandikota, J. Materzyńska, J. Fiotto-Kaufman, and D. Bau, “Erasing Concepts from Diffusion Models,” in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2426–2436, IEEE, 2023.

[30] R. Eldan and M. Russinovich, “Who’s harry potter? approximate unlearning in llms,” in *International Conference on Learning Representations (ICLR)*, 2024.

[31] F. Huang *et al.*, “Catastrophic jailbreak of open-source llms via exploiting generation,” 2023.

[32] L. Struppek *et al.*, “Rickrolling the reader: Injecting invisible backdoors into text encoders,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[33] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, “Training language models to follow instructions with human feedback,” in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS ’22, Curran Associates Inc., 2022.

[34] R. Rafailov *et al.*, “Direct preference optimization: Your language model is secretly a reward model,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[35] G. Seneviratne, J. An, S. Ellahy, K. Weerakoon, M. B. Elnoor, J. D. Kannan, A. T. Sunil, and D. Manocha, “Halo: Human preference aligned offline reward learning for robot navigation,” 2025.

[36] Y. Wolf *et al.*, "Fundamental limits of alignment in large language models," in *International Conference on Learning Representations (ICLR)*, 2024.

[37] B. Wallace *et al.*, "Universal adversarial triggers for vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[38] Y. Liu *et al.*, "Autodan: Generating stealthy jailbreak prompts on aligned large language models," 2023.

[39] J. Rando *et al.*, "Universal jailbreak backdoors from poisoned human feedback," 2023.

[40] R. Debenedetti *et al.*, "Privacy side channels in large language models," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2023.

[41] J. Kirchenbauer *et al.*, "A watermark for large language models," in *International Conference on Machine Learning (ICML)*, 2023.

[42] Z. Yuan *et al.*, "Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher," in *International Conference on Learning Representations (ICLR)*, 2024.

[43] M. Nasr *et al.*, "Scalable extraction of training data from (production) language models," 2023.

[44] N. Carlini *et al.*, "Stealing part of a production language model," 2024.

[45] M. Andriushchenko *et al.*, "Jailbreaking is not enough: The need for holistic safety," in *International Conference on Machine Learning (ICML)*, 2024.

[46] M. Mazeika, L. Phan, X. Yin, A. Zou, Z. Wang, N. Mu, E. Sakhaee, N. Li, S. Basart, B. Li, D. Forsyth, and D. Hendrycks, "Harmbench: a standardized evaluation framework for automated red teaming and robust refusal," in *Proceedings of the 41st International Conference on Machine Learning*, ICML'24, JMLR.org, 2024.

[47] S.-Y. Chen *et al.*, "Trojdiff: Trojan attacks on diffusion models with diverse targets," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[48] N. R. Haddaway, M. J. Page, C. C. Pritchard, and L. A. McGuinness, "Prisma2020: An r package and shiny app for producing prisma 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis," *Campbell Systematic Reviews*, vol. 18, no. 2, p. e1230, 2022.

[49] D. Hendrycks, M. Mazeika, A. Zou, S. Patel, C. Zhu, J. Navarro, D. Song, B. Li, and J. Steinhardt, "What would jiminy cricket do? towards agents that behave morally," in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

[50] S. Zhu *et al.*, "Autodan: Automatic and interpretable adversarial attacks on large language models," 2023.

[51] R. Lapid *et al.*, "Open sesame: Universal black box jailbreaking of large language models," in *IEEE International Conference on Communications (ICC)*, 2024.

[52] J.-Y. Jiang *et al.*, "Artprompt: Ascii art jailbreak against aligned llms," in *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

[53] Anthropic Research, "Many-shot jailbreaking." Technical Report, 2024.

[54] F. Li *et al.*, "Deepinception: Hypnotizing large language models into long-term deception," 2023.

[55] N. Xu, B. Gao, and H. Dou, "Jailbreaking llms via semantically relevant nested scenarios with targeted toxic knowledge," 2025.

[56] N. Zverev *et al.*, "Multi-turn jailbreak: Breaking alignment with conversation," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[57] H. Wang *et al.*, "Foot in the door: Accumulative jailbreak attacks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[58] H. Yoo, Y. Yang, and H. Lee, "Code-switching red-teaming: LLM evaluation for safety and multilingual understanding," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13392–13413, Association for Computational Linguistics, 2025.

[59] C. Sitawarin *et al.*, "Pre-training poisoning of llms," 2024.

[60] Y. Zou *et al.*, "Poisoned rag: Knowledge injection attacks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[61] R. Schaeffer, D. Valentine, L. Bailey, J. Chua, C. Eyzaguirre, Z. Durante, J. Benton, B. Miranda, H. Sleight, T. T. Wang, J. Hughes, R. Agrawal, M. Sharma, S. Emmons, S. Koyejo, and E. Perez, "Failures to find transferable image jailbreaks between vision-language models," in *International Conference on Learning Representations*, ICLR, 2025.

[62] J. Broomfield, G. Ingebretsen, R. Iranmanesh, S. Pieri, E. Kosak-Hine, T. Gibbs, R. Rabbany, and K. Pelrine, "Decompose, recompose, and conquer: Multi-modal LLMs are vulnerable to compositional adversarial attacks in multi-image queries," in *Red Teaming GenAI: What Can We Learn from Adversaries?*, 2025.

[63] Y. Fan *et al.*, "Agent-smith: A single image can jailbreak one million multimodal llms," in *International Conference on Machine Learning (ICML)*, 2024.

[64] G.-H. Chou *et al.*, "Baddiffusion: How to backdoor diffusion models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[65] W. Lin *et al.*, "Backdoordm: A comprehensive benchmark for backdoor learning in diffusion model," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[66] S. Shan *et al.*, "Glaze: Protecting artists from style mimicry by glazing," in *USENIX Security Symposium*, 2023.

[67] Y. Wen, J. Kirchenbauer, J. Geiping, and T. Goldstein, "Tree-rings watermarks: invisible fingerprints for diffusion images," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Curran Associates Inc., 2023.

[68] J. Liang, Z. Wang, S. Hong, S. Ji, and T. Wang, "Watermark under fire: A robustness evaluation of llm watermarking," in *The 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*, pp. 21050–21074, 01 2025.

[69] V. Sadasivan *et al.*, "Can ai-generated text be reliably detected?," 2023.

[70] A. Zhou, B. Li, and H. Wang, "Robust prompt optimization for defending language models against jailbreaking attacks," in *Advances in Neural Information Processing Systems*, vol. 37, Curran Associates, Inc., 2024.

[71] Y. Xie, J. Yi, J. Shao, J. Curl, L. Lyu, Q. Chen, X. Xie, and F. Wu, "Defending chatgpt against jailbreak attack via self-reminders," *Nature Machine Intelligence*, vol. 5, pp. 1486–1496, 2023.

[72] S. Xhonneux, A. Sordoni, S. Günnemann, G. Gidel, and L. Schwinn, "Efficient adversarial training in llms with continuous attacks," in *Advances in Neural Information Processing Systems*, vol. 37, pp. 1502–1530, Curran Associates, Inc., 2024.

[73] G. Li *et al.*, "Tensortrust: System-level security for llms," in *Symposium on Operating Systems Design and Implementation (OSDI)*, 2024.

[74] G. Bansal *et al.*, "Cleanclip: Mitigating data poisoning attacks in multimodal contrastive learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[75] G. Wang *et al.*, "Decodingtrust: A comprehensive assessment of trustworthiness in gpt models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[76] S. Bhatt *et al.*, "Purple teaming: Automated adversarial evaluation," 2024.

[77] J. Zhou, K. Li, J. Li, J. Kang, M. Hu, X. Wu, and H. Meng, "Purple-teaming llms with adversarial defender training," 2024.

[78] B. Y. Lin, R. L. Bras, K. Richardson, A. Sabharwal, R. Poovendran, P. Clark, and Y. Choi, "Zebralogic: On the scaling limits of llms for logical reasoning," 2025.

[79] M. Mazeika, D. Hendrycks, H. Li, X. Xu, S. Hough, A. Zou, A. Rajabi, Q. Yao, Z. Wang, J. Tian, Y. Tang, D. Tang, R. Smirnov, P. Pleskov, N. Benkovich, D. Song, R. Poovendran, B. Li, and D. Forsyth, "The trojan detection challenge," in *Proceedings of the NeurIPS 2022 Competitions Track*, vol. 220 of *Proceedings of Machine Learning Research*, pp. 279–291, PMLR, 28 Nov–09 Dec 2022.

[80] Y. Shu *et al.*, "On the exploitability of instruction tuning," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[81] Z. Liang *et al.*, "Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples," in *International Conference on Machine Learning (ICML)*, 2023.

[82] H. Salman *et al.*, "Raising the cost of malicious ai-powered image editing," in *International Conference on Machine Learning (ICML)*, 2023.

[83] Z. Dai *et al.*, "Safe rlhf: Constrained reinforcement learning for safety," in *International Conference on Learning Representations (ICLR)*, 2024.

[84] Y. Zhou *et al.*, "Fortify the shortest stave: Towards robust llm safety," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[85] K. Li *et al.*, "Inference-time intervention: Eliciting truthfulness from llms," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[86] C. Burns *et al.*, "Discovering latent knowledge in language models without supervision," in *International Conference on Learning Representations (ICLR)*, 2023.

[87] M. Mazeika, B. Li, and D. Forsyth, "How to steer your adversary: Targeted and efficient model stealing defenses with gradient redirection," in *International Conference on Machine Learning*, 2022.

[88] H. Huang *et al.*, "Trustllm: Trustworthiness in large language models," 2024.

[89] Y. Li *et al.*, "A whac-a-mole game: Bypass and defense of llm safety," 2024.

[90] M. Mazeika, E. Tang, A. Zou, S. Basart, J. S. Chan, D. Song, D. Forsyth, J. Steinhardt, and D. Hendrycks, “How would the viewer feel? estimating wellbeing from video scenarios,” in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS ’22, Curran Associates Inc., 2022.