A Survey on Attack Defend and Evaluate in AI Security with Focus on Prompt Engineering and Adversarial Attacks

www.surveyx.cn

Abstract

This survey paper explores a comprehensive framework for AI security, focusing on the interplay between adversarial attacks, defenses, and prompt engineering. It addresses the vulnerabilities of AI systems, particularly deep neural networks, to adversarial attacks, which exploit model weaknesses to manipulate outcomes. The paper emphasizes the significance of AI security in critical applications, where adversarial manipulation can compromise system integrity. Defensive strategies, including adversarial training, dropout, noise-based defenses, and GAN-based approaches, are analyzed for their effectiveness in enhancing model robustness. The role of prompt engineering is highlighted as a pivotal technique in mitigating adversarial threats by strategically crafting inputs to improve AI performance and resilience. The survey also introduces innovative frameworks and benchmarks for evaluating AI security, providing a systematic approach to assess model robustness against adversarial attacks. Through an in-depth analysis of recent advancements in adversarial attack strategies and defensive mechanisms, this paper contributes to the ongoing discourse on AI security by offering insights into the evolving landscape of adversarial threats and defenses. Ultimately, it aims to guide future developments in securing AI systems, ensuring their robustness and reliability in the face of emerging adversarial challenges.

1 Introduction

1.1 Significance of AI Security

The integration of artificial intelligence (AI) into various sectors has notably enhanced capabilities while simultaneously introducing significant security challenges. Vulnerabilities in machine learning algorithms can be exploited by adversarial examples, which mislead predictions and highlight the critical importance of AI security [1]. The sophistication of cyber threats necessitates robust defenses for AI systems, particularly in critical infrastructures [2]. The susceptibility of deep neural networks (DNNs) to adversarial attacks poses considerable risks, as adversaries can manipulate outcomes, especially in sensitive applications like biometric authentication and network intrusion detection [3, 4].

Machine learning models, foundational to many AI applications, are particularly vulnerable to adversarial attacks and data poisoning, underscoring the urgent need for effective AI security measures [5]. The complexity of modern cyber-attacks, combined with the operational risks posed by these vulnerabilities, further accentuates the necessity of securing AI systems [6]. In cyber-physical systems (CPS), adversary emulation introduces additional challenges, as these systems operate in environments where precision and reliability are crucial [7].

The incorporation of AI in industrial environments also exposes systems to novel security risks that demand high precision and reliability [8]. The disparity between the rapid proliferation of

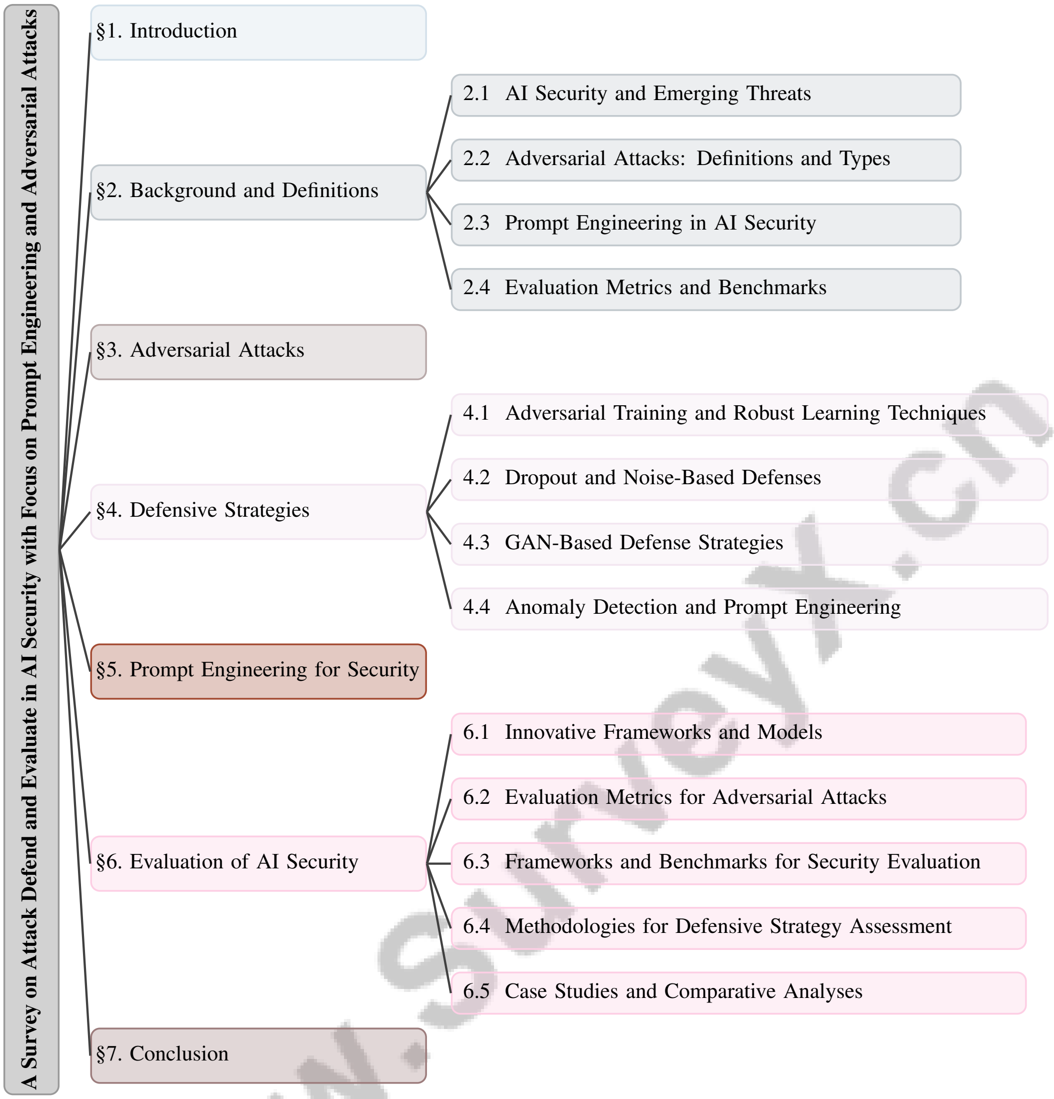

A Survey on Attack Defend and Evaluate in AI Security with Focus on Prompt Engineering and Adversarial Attacks
§1. Introduction
§2. Background and Definitions
2.1 AI Security and Emerging Threats
2.2 Adversarial Attacks: Definitions and Types
2.3 Prompt Engineering in AI Security
2.4 Evaluation Metrics and Benchmarks
§3. Adversarial Attacks
§4. Defensive Strategies
4.1 Adversarial Training and Robust Learning Techniques
4.2 Dropout and Noise-Based Defenses
4.3 GAN-Based Defense Strategies
4.4 Anomaly Detection and Prompt Engineering
§5. Prompt Engineering for Security
§6. Evaluation of AI Security
6.1 Innovative Frameworks and Models
6.2 Evaluation Metrics for Adversarial Attacks
6.3 Frameworks and Benchmarks for Security Evaluation
6.4 Methodologies for Defensive Strategy Assessment
6.5 Case Studies and Comparative Analyses
§7. Conclusion

Figure 1: chapter structure

connected devices and the shortage of cybersecurity professionals highlights the need for robust frameworks to secure machine learning models against adversarial threats [9]. Furthermore, the ethical implications of AI vulnerabilities, particularly those associated with large language models (LLMs) in contexts such as surveillance and copyright theft, illuminate broader societal risks stemming from inadequate AI security [10]. Penetration testing remains a reliable method for assessing network security, ensuring AI systems are resilient against potential attacks. As AI continues to permeate various domains, the significance of its security is paramount, necessitating ongoing research and development of advanced security measures.

1.2 Role of Adversarial Attacks and Defenses

Adversarial attacks pose a substantial challenge to the robustness and reliability of AI systems by exploiting vulnerabilities in machine learning models, particularly DNNs, through subtle perturbations often undetectable by human observers. These perturbations can result in critical misclassifications in applications such as autonomous vehicles and surveillance systems, where adversarial attacks can compromise detection methods like Non-Maximum Suppression (NMS) [11]. The transferability of adversarial examples across different model architectures complicates the formulation of universal defense strategies [12].

In language models, adversarial attacks, including backdoor attacks, introduce significant risks by embedding malicious triggers that manipulate outputs, creating challenges in maintaining the integrity of LLMs [13]. Adversarial word substitution further exemplifies these challenges, misleading pre-trained language models (PrLMs) into erroneous predictions, necessitating robust defense mechanisms to ensure model reliability [14].

Defensive strategies are essential for mitigating the impact of adversarial attacks, with adversarial training emerging as a key approach. This method integrates adversarial examples into the training process, enhancing model resilience to perturbations [2]. However, the adaptability of traditional defenses is often limited, particularly against advanced, targeted attacks that require extensive retraining [15]. The dual nature of adversarial attacks, encompassing both white-box and black-box strategies, necessitates comprehensive and dynamic defenses. White-box attacks exploit full knowledge of the target model to craft precise adversarial examples, while black-box attacks leverage transferability and querying strategies without direct model access [3].

The significance of adversarial defenses in AI security is underscored by their potential to prevent operational failures, such as misclassification in power quality recognition tasks that could disrupt electrical grids [3]. As adversarial attacks and defenses evolve, securing AI systems requires a dynamic and adaptive approach to both attack and defense mechanisms, ensuring resilience against emerging threats.

1.3 Objectives and Significance of the Paper

This survey aims to provide a comprehensive analysis of adversarial attacks and corresponding defense mechanisms within the domain of AI security, particularly focusing on prompt engineering. It explores vulnerabilities in neural networks exploited by adversarial attacks and proposes robust defense mechanisms to enhance AI systems' resilience against such threats [3]. By investigating the impact of prompt injection attacks and assessing the effectiveness of secure prompting strategies, this study seeks to improve both system performance and security [2].

In addition to addressing vulnerabilities, the survey defines comprehensive attack models for machine learning systems within the context of cybersecurity, emphasizing operational robustness in a defense-in-depth environment. This involves developing effective countermeasures against evolving cyber threats by analyzing attacks from the attacker's perspective [16]. A significant contribution of this survey is the introduction of a model for probabilistic attack planning that incorporates uncertainty, crucial for devising more effective attack strategies.

Furthermore, the survey highlights novel methods for detecting backdoor attacks using adversarial examples, thereby contributing to AI security enhancement. It also aims to advance the development of robust defense strategies against gradient-based attack techniques, such as BPDA and EOT, providing a benchmark for real-time comparison of different adversarial attack and defense techniques [3].

Through these objectives, this paper aspires to significantly contribute to AI security research by offering insights into the evolving landscape of adversarial attacks and defenses. The ultimate goal is to establish a comprehensive framework for the future development of AI systems focused on enhancing cybersecurity, ensuring robustness and reliability against increasingly sophisticated adversarial threats through advanced AI-driven solutions for threat detection, vulnerability assessment, and incident response. Additionally, it emphasizes the importance of addressing ethical considerations and technical limitations inherent in AI applications, fostering collaboration among policymakers, organizations, and cybersecurity practitioners to navigate the evolving landscape of cyber threats [17, 18, 19].

1.4 Structure of the Survey

This survey is meticulously structured to provide a comprehensive understanding of the complex domain of AI security, with a focus on adversarial attacks and prompt engineering. It critically examines the role of adversarial perturbations in natural language processing (NLP) security, evaluating the effectiveness of current methodologies against real-world threats, and highlights the necessity for a systematic approach to model stealing attacks and defenses in Machine Learning-as-a-Service (MLaaS). By integrating insights from recent research, this survey clarifies problem definitions, research goals, and practical implications for enhancing AI system security [20, 19]. The paper begins with an introduction that establishes the significance of AI security, emphasizing the critical role of adversarial attacks and defenses in safeguarding AI systems. Following this, the objectives and significance of the survey are outlined, setting the stage for a detailed exploration of the topic.

The second section, Background and Definitions, offers foundational knowledge on key concepts such as AI security, adversarial attacks, and prompt engineering, defining evaluation metrics essential for assessing AI security and equipping readers with necessary context for subsequent discussions.

In the third section, Adversarial Attacks, the survey delves into various types of adversarial attacks, including white-box and black-box attacks, and examines techniques for generating adversarial examples. This section reviews recent advancements in adversarial attack strategies, providing a thorough understanding of the current threat landscape.

The fourth section, Defensive Strategies, investigates methods employed to defend AI systems against adversarial attacks, including adversarial training, dropout, noise-based defenses, and GAN-based strategies. It also discusses the roles of anomaly detection and prompt engineering in enhancing defenses.

Prompt Engineering for Security, the fifth section, explores how prompt engineering can improve AI performance and resilience, reviewing case studies and examples where prompt engineering has been effectively applied, emphasizing its potential in cybersecurity.

In the sixth section, Evaluation of AI Security, the survey discusses metrics and methodologies used to evaluate AI systems' robustness. It explores innovative frameworks and benchmarks that assess the effectiveness of defensive strategies and the resilience of AI models against adversarial attacks.

Finally, the Conclusion summarizes key findings and insights, highlighting ongoing challenges and future directions in AI security. By organizing current research into fields such as threat detection, incident response, and proactive defense mechanisms, the survey provides a holistic view of the landscape, utilizing criteria such as effectiveness and ethical considerations [18].The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 AI Security and Emerging Threats

AI security is pivotal in protecting systems from threats, particularly adversarial attacks exploiting machine learning model vulnerabilities. These attacks often involve slight data alterations, causing significant misclassification errors with severe consequences in critical applications such as autonomous driving, where misreading traffic signs can be life-threatening [21]. Current evaluation methods for pattern classifiers in adversarial settings are inadequate for handling non-stationary data, necessitating robust frameworks [22].

Emerging threats in AI security are growing in complexity. Large language models (LLMs) are susceptible to backdoor attacks embedding undetectable malicious triggers, compromising AI system integrity [13]. Additionally, the power imbalance between data owners and LLMs facilitates harmful inferences without consent, complicating the security landscape [10]. Advanced attack vectors like model hijacking and deployment-stage backdoor attacks further complicate defenses by embedding backdoors in models with minimal access to parameters, challenging existing methods [9].

AI integration into 5G networks has transformed the security landscape, introducing sophisticated cyber threats leveraging AI capabilities [23]. Adversarial manipulation of reward feedback in meta-reinforcement learning poses another threat, leading agents to learn incorrect policies, compromising decision-making [24]. As AI becomes integral to critical infrastructures, addressing these threats is crucial. Traditional Cyber Threat Intelligence (CTI) methods often fall short against sophisticated threats [25]. Deep learning models for malware detection remain vulnerable to adversarial malware evading detection while retaining malicious functionality [26]. These challenges underscore the urgent need for advanced security measures to protect AI systems.

2.2 Adversarial Attacks: Definitions and Types

Adversarial attacks manipulate input data to produce erroneous outputs, often undetectable to human observers. These attacks exploit AI system vulnerabilities, particularly in deep neural networks

(DNNs), by introducing subtle perturbations that degrade performance, notably in image classification [15]. Adversarial attacks are primarily categorized into white-box and black-box attacks. White-box attacks assume full access to the model's architecture and parameters, enabling precise adversarial example crafting to maximize misclassification [12]. Black-box attacks operate with limited or no access to the model's internal details, relying on queries to infer behavior and generate adversarial examples [27].

Specialized forms include signal-specific adversarial attacks (SSA) that generate minimal perturbations tailored to specific signals, and signal-agnostic adversarial attacks (SAA) creating universal perturbations for multiple signals [3]. Backdoor attacks involve poisoned training data implanting backdoors in models, causing incorrect predictions when specific triggers are present [13]. The complexity of adversarial attacks is further highlighted by transferable adversarial examples affecting multiple models simultaneously [12], complicating defense strategies and necessitating robust models against both direct attacks and examples crafted for other models.

Understanding adversarial attack definitions and types is vital for developing effective defenses and ensuring AI system robustness. As adversarial techniques evolve, addressing challenges in practical applications is critical, particularly in natural language processing (NLP), where adversarial samples significantly impact performance and security. Current research emphasizes security-oriented approaches considering real-world attacker motivations and developing defense mechanisms capable of detecting and mitigating these attacks, ensuring AI system resilience. Understanding adversarial perturbations in images or text is vital for creating effective defenses, highlighting the necessity for collaboration among researchers, practitioners, and policymakers to enhance AI security protocols [17, 28, 19, 29, 30].

2.3 Prompt Engineering in AI Security

Prompt engineering is crucial in enhancing AI security by designing input prompts to mitigate adversarial threats. This approach is especially significant for large language models (LLMs), where adversarial prompt injections manipulate outputs by exploiting instruction-following capabilities [10]. Developing robust defenses against such vulnerabilities is essential as these models are increasingly integrated into high-security applications.

Prompt engineering plays a vital role in adversarial frameworks like TextAttack, enabling model-agnostic and dataset-agnostic adversarial attacks, allowing users to evaluate AI system resilience without extensive code modifications [31]. This flexibility is crucial for testing various models under adversarial conditions, enhancing overall security posture. Moreover, prompt engineering is instrumental in creating benchmarks emphasizing secure prompting and assessing mitigation strategies against adversarial inputs. These benchmarks provide insights into AI system resilience, differing from traditional ones by focusing specifically on security aspects [31].

In offensive strategies, prompt engineering generates adversarial examples resembling original images but misclassified by face recognition networks, demonstrating adversarial prompts' impact in testing AI system robustness [32]. The integration of prompt engineering in tools like Adagio allows interactive experimentation with adversarial attacks and defenses on Automatic Speech Recognition (ASR) models, incorporating real-time audio processing techniques [7]. This capability is vital for evaluating and enhancing AI system robustness in dynamic environments.

Prompt engineering also improves detection systems in industrial control systems (ICS) by providing a structured approach to evaluate and enhance detection mechanisms, emphasizing its role in bolstering security measures [2]. The strategic use of prompt engineering in generating adversarial prompts, enhancing model resilience, and facilitating strategic decision-making underscores its significance in the evolving AI security landscape.

As AI technologies advance, prompt engineering's role in ensuring resilience and reliability against emerging cybersecurity challenges is paramount. Ongoing advancements and applications of prompt engineering techniques are crucial for developing and sustaining resilient AI security frameworks, defending against increasingly sophisticated adversarial threats, such as prompt injection attacks manipulating AI responses and compromising system integrity. This necessitates a comprehensive understanding of security vulnerabilities and implementing automated defenses and robust testing methodologies to enhance AI system effectiveness in real-world applications [17, 33, 34, 35].

2.4 Evaluation Metrics and Benchmarks

Benchmark	Size	Domain	Task Format	Metric
Multi-SpacePhish[36]	4,000	Phishing Website Detection	Adversarial Attack Evaluation	True Positive Rate, False Positive Rate
ALP[37]	1,000,000	Image Classification	Adversarial Example Evaluation	Adversarial Accuracy
RMC[38]	71	Computer Vision	Classification	ECE, Accuracy
AAB[39]	70,000	Image Classification	Digit Recognition	Accuracy, Adversarial Accuracy
MEMBERSHIP-DOCTOR[40]	60,000	Image Classification	Membership Inference	Accuracy, Attack Success Rate
SAM-Bench[41]	200	Image Segmentation	Adversarial Attack Evaluation	mIoU
DUMB[42]	13,000	Computer Vision	Binary Classification	F1 Score
CodeAttack[43]	46,680	Code Translation	Code-Code	CodeBLEU, BLEU

Table 1: This table presents a comprehensive overview of various benchmarks employed to evaluate AI security against adversarial attacks. It includes details on benchmark size, domain, task format, and metrics used, providing a structured comparison of different evaluation frameworks. The benchmarks cover a range of applications, from phishing website detection to image classification and code translation, highlighting the diversity of tasks in AI security assessment.

Evaluating AI security, particularly against adversarial attacks and defenses, relies on robust metrics and benchmarks accurately assessing model resilience and effectiveness. These metrics identify vulnerabilities in deep neural networks and evaluate defense mechanisms against adversarial threats. Establishing a quantitative framework for assessing a model's intrinsic robustness enables comprehensive analysis of neural network stability under adversarial attacks. Moreover, these metrics facilitate identifying misclassified regions and enhance detecting unseen adversarial samples, improving overall robustness and detection efficiency [44, 19, 30, 45, 46].

Table 1 provides a detailed overview of representative benchmarks used in evaluating AI security, illustrating the diversity in domains, task formats, and metrics applied in adversarial attack assessments. Common metrics such as accuracy and BLEU score evaluate model performance regarding robustness against adversarial attacks, serving as benchmarks for assessing AI security [47]. The Empirical Evaluation Framework for Classifier Security (EECS) offers a systematic approach to evaluating classifier security by simulating attack scenarios and modeling adversaries, providing a comprehensive assessment of model robustness [22].

The critical-time metric enhances risk analysis and treatment for controlled systems by providing a time-based measure for safety duration following an anomaly, valuable in applications requiring real-time decision-making [48]. This metric is complemented by the True Positive Rate (TPR) and False Positive Rate (FPR), assessing model performance in detecting adversarial threats, such as phishing websites [36].

In cyber-physical systems, systematically breaking down the seven phases of the Cyber Attack Thread offers an evaluation method assessing defenses' effectiveness at each stage, providing a holistic view of system security [49]. This comprehensive approach underscores the importance of evaluating AI security across multiple dimensions, ensuring defenses are robust against various attack vectors.

Improved data collection practices, particularly regarding timing-related data, are emphasized as essential for evaluating AI security. Such practices are crucial for developing accurate and reliable security metrics reflecting true adversarial robustness [4]. These practices ensure evaluation metrics provide a comprehensive understanding of the model's security posture, guiding the development of more resilient AI systems.

3 Adversarial Attacks

Adversarial attacks exploit vulnerabilities in AI systems, challenging their robustness and reliability. Understanding the methods for generating adversarial examples is crucial for developing defenses against these sophisticated strategies, which can significantly impact AI security. As illustrated in Figure 2, the hierarchical structure of adversarial attacks categorizes various techniques for generating adversarial examples and highlights their impact on AI models. This figure delineates gradient-based, decision-based, and NLP-specific techniques, while also discussing critical vulnerabilities and transferability issues. Additionally, it showcases recent advancements in methods that enhance

AI-generated, for reference only.

Category	Feature	Method
Techniques for Generating Adversarial Examples	Model Robustness Strategies	DI2-FGSM[50], GOT[51]
	Gradient-Free Methods	XSUB[27]
	Adversarial Detection and Deception	NNIF[52]
	Optimization Techniques	MGS[53], GF[54], PD[55]
Impact on AI Models	Adversarial Perturbation Techniques	BA[56], OPA[57], GBDA[58]
	Attack Complexity Strategies	SS[59]
	Cross-Model Threats	HA[60], EAEG[12]
Recent Advancements in Adversarial Attack Strategies	Optimization Techniques	AW[61]

Table 2: This table provides a comprehensive summary of various methods and strategies employed in the generation of adversarial examples and their impact on AI models. It categorizes techniques into model robustness strategies, gradient-free methods, adversarial detection and deception, and optimization techniques, highlighting recent advancements in adversarial attack strategies. The table serves as a valuable resource for understanding the evolving landscape of adversarial attacks and their implications for AI security.

both the success and robustness of these attacks, thereby providing a comprehensive overview of the evolving landscape of adversarial strategies. Table 2 presents a detailed summary of the techniques for generating adversarial examples, their impact on AI models, and recent advancements in adversarial attack strategies, offering insights into the complexity and sophistication of these methods.

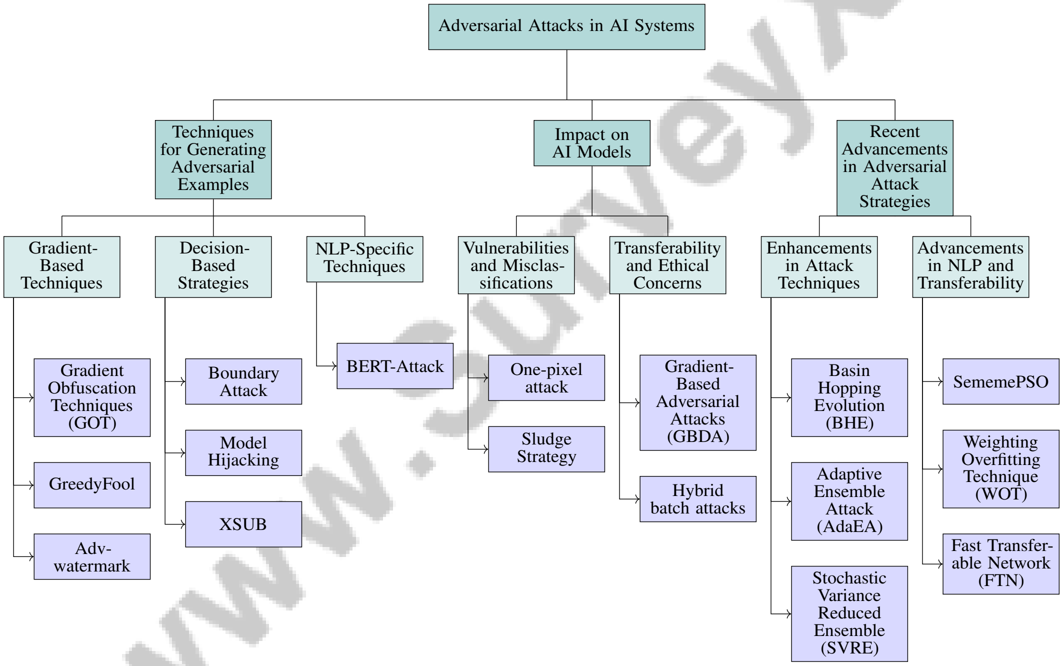

Adversarial Attacks in AI Systems
Techniques for Generating Adversarial Examples
Impact on AI Models
Recent Advancements in Adversarial Attack Strategies
Gradient-Based Techniques
Decision-Based Strategies
NLP-Specific Techniques
Vulnerabilities and Misclassifications
Transferability and Ethical Concerns
Enhancements in Attack Techniques
Advancements in NLP and Transferability
Gradient Obfuscation Techniques (GOT)
GreedyFool
Adv-watermark
Boundary Attack
Model Hijacking
XSUB
BERT-Attack
One-pixel attack
Sludge Strategy
Gradient-Based Adversarial Attacks (GBDA)
Hybrid batch attacks
Basin Hopping Evolution (BHE)
Adaptive Ensemble Attack (AdaEA)
Stochastic Variance Reduced Ensemble (SVRE)
SememePSO
Weighting Overfitting Technique (WOT)
Fast Transferable Network (FTN)

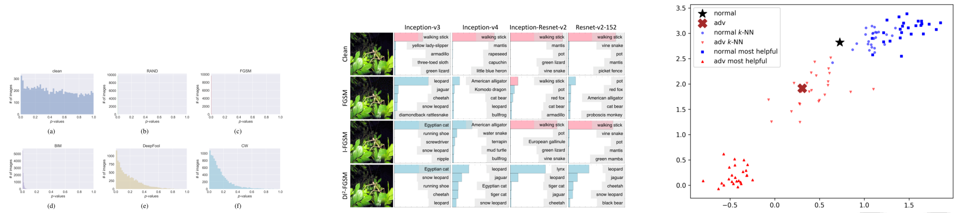
Figure 2: This figure illustrates the hierarchical structure of adversarial attacks in AI systems, categorizing techniques for generating adversarial examples, their impact on AI models, and recent advancements in attack strategies. It highlights gradient-based, decision-based, and NLP-specific techniques, discusses vulnerabilities and transferability issues, and showcases new methods enhancing attack success and robustness.

3.1 Techniques for Generating Adversarial Examples

Creating adversarial examples involves manipulating inputs to deceive machine learning models, particularly deep neural networks (DNNs), into incorrect predictions. Gradient Obfuscation Techniques (GOT) mitigate attacks by disrupting gradient-based optimizations [51]. GreedyFool uses differential evolution to generate imperceptible perturbations, maintaining visual fidelity while deceiving models [54]. Adv-watermark embeds watermarks in images to mislead models without compromising visual integrity [61]. The Boundary Attack reduces perturbations iteratively, demonstrating decision-based strategies' effectiveness without gradient reliance [56].

In cybersecurity, model hijacking attacks like MalGuise manipulate control-flow graphs to evade detection [53]. XSUB misleads black-box classifiers in text-based applications by substituting key features [27]. Ensemble-based Adversarial Example Generation (EAEG) enhances example transferability across models, challenging defenses reliant on model-specific traits [12].

The complexity of crafting adversarial examples in natural language processing (NLP) remains a significant challenge due to text data's discrete nature. Techniques like BERT-Attack maintain semantic integrity while generating effective adversarial samples, emphasizing the need for security-oriented research aligned with real-world scenarios [29, 19].



(a) Visualization of p-values for different attack methods on CIFAR-10 dataset[55]

(b) The image shows a comparison of different models' predictions on a set of images, with the models being Inception-v3, Inception-v4, Inception-ResNet-v2, and ResNet-v2-152. The models are tested on a dataset containing various animals and objects, and the predictions are visualized in a bar chart format.[50]

(c) The image shows a scatter plot with various data points representing different classes and their corresponding features.[52]

Figure 3: Examples of Techniques for Generating Adversarial Examples

Figure 3 illustrates various adversarial attack techniques. The first subfigure shows p-values for different attacks on the CIFAR-10 dataset, highlighting method efficacy. The second subfigure compares model predictions under adversarial conditions, indicating susceptibility variations. The third subfigure categorizes data points by class and features, emphasizing the diverse strategies and implications of adversarial attacks [55, 50, 52].

3.2 Impact on AI Models

Adversarial attacks threaten AI models by exploiting vulnerabilities, leading to erroneous outputs with severe implications. Subtle perturbations, like the one-pixel attack, achieve high misclassification rates across DNNs [57]. The Sludge Strategy imposes operational challenges, highlighting adversarial techniques' potential to disrupt AI systems [59].

Figure 4 illustrates the impact of adversarial attacks on AI models, categorizing them into different types of attacks, strategies, and defense mechanisms. This figure not only highlights the vulnerabilities of AI systems to various attack methods but also underscores the pressing need for robust defenses.

In language models, Gradient-Based Adversarial Attacks (GBDA) preserve fluency while reducing transformer model accuracy [58]. Hybrid batch attacks improve black-box attack efficiency, exemplifying evolving strategies [60]. Decision-based attacks like the Boundary Attack maintain effectiveness with fewer queries [56].

Adversarial examples' transferability complicates defenses, necessitating robust models against examples crafted for other models [12]. Attacks reveal AI systems' limitations, including biases and ethical dilemmas, necessitating robust risk management [18, 48]. Continuous advancements in defense mechanisms are essential to safeguard AI systems against these evolving threats [38, 29, 28].

3.3 Recent Advancements in Adversarial Attack Strategies

Recent advancements in adversarial attack strategies enhance attackers' ability to deceive AI models. Basin Hopping Evolution (BHE) improves attack success rates in black-box settings [61]. In computer

AI-generated, for reference only.

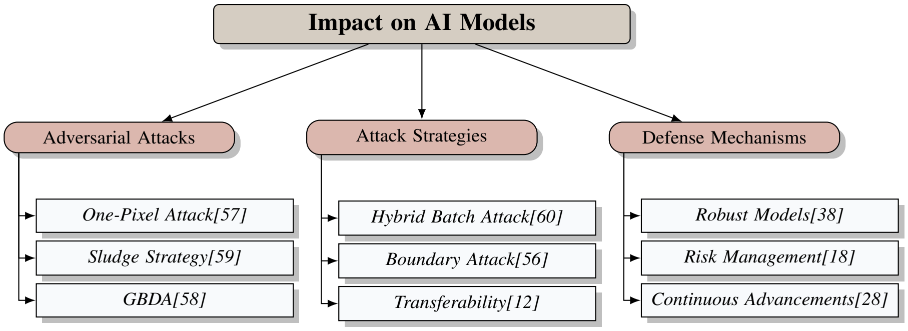


Figure 4: This figure illustrates the impact of adversarial attacks on AI models, categorizing them into different types of attacks, strategies, and defense mechanisms. It highlights the vulnerabilities of AI systems to various attack methods and underscores the need for robust defenses.

vision, experiments on ImageNet with models like Inception v3 and v4 reveal DNNs' susceptibility to adversarial examples, prompting advancements in adversarial training and detection methods [38, 62, 19, 55, 63].

Ensemble-based approaches, such as Adaptive Ensemble Attack (AdaEA) and Stochastic Variance Reduced Ensemble (SVRE), enhance adversarial example transferability across models, optimizing ensemble processes for effectiveness [64, 65, 66, 67]. In NLP, innovations like SememePSO improve model robustness through adversarial training [28, 19, 29, 68, 69].

Techniques like the Weighting Overfitting Technique (WOT) and Fast Transferable Network (FTN) enhance adversarial robustness and transferability across classifiers [70, 71, 72]. These advancements underscore the need for continuous innovation in defense mechanisms to address sophisticated threats, emphasizing collaboration among policymakers, cybersecurity practitioners, and researchers to enhance digital infrastructure resilience [17, 18, 19, 29, 69].

4 Defensive Strategies

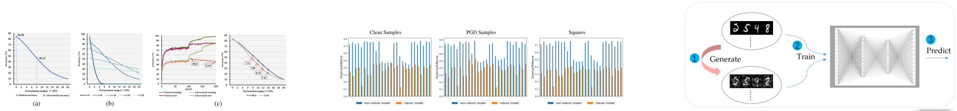
Developing robust defenses against adversarial attacks is crucial for enhancing AI systems' resilience. This section delves into foundational techniques aimed at fortifying models against such threats, emphasizing adversarial training and robust learning techniques. These approaches not only bolster model performance under attack but also contribute significantly to AI security discourse. Table 4 provides a comparative analysis of various defensive strategies utilized in AI systems to mitigate adversarial attacks, detailing their primary mechanisms, defense focus, and unique features. The subsequent subsection elaborates on adversarial training's significance, detailing its mechanisms and impact on model robustness.

4.1 Adversarial Training and Robust Learning Techniques

Adversarial training enhances AI model robustness by integrating adversarial examples into the training process, enabling models to withstand perturbations effectively [3]. A notable advancement is metric learning approaches like TLA, optimizing distances between clean and adversarial representations to strengthen defenses [45]. The tunable security framework aligns with adversarial training by dynamically adjusting security measures to balance protection levels and costs [2]. FM-Defense exemplifies an attack-agnostic method that purifies adversarial examples by manipulating semantic features through a combo-variational autoencoder, showcasing semantic-based defenses' potential [15].

Beyond adversarial training, robust learning techniques such as ensemble-based methods enhance the transferability of targeted adversarial examples, improving model robustness across diverse architectures [12]. These methods highlight the necessity for comprehensive defenses capable of addressing various adversarial strategies. Techniques like XSUB demonstrate adaptability with high effectiveness under minimal perturbations, crucial for real-world applications facing varied adversarial scenarios, including backdoor attacks [27]. However, current defenses primarily target data poisoning and may falter against architectural backdoor attacks, indicating a need for further innovation [13].

Adversarial training and robust learning techniques are pivotal in developing resilient AI systems. Continuous advancement of AI methodologies, coupled with innovative strategies and thorough analyses, is vital for enhancing AI technologies' security and reliability amid increasingly sophisticated cyber threats. This dynamic landscape necessitates a multifaceted cybersecurity approach, where AI plays a crucial role in threat detection, incident response, and proactive risk mitigation. As organizations navigate complex adversarial environments, addressing both technical challenges and ethical considerations associated with AI deployment is imperative [17, 18].



(a) Comparison of Natural and Adversarial Training and Testing Accuracy for Different Perturbation Budgets[73]

(b) Comparison of Overconfidence Between Non-Robust and Robust Models on Different Types of Data[38]

(c) Deep Learning for Image Generation and Prediction[74]

Figure 5: Examples of Adversarial Training and Robust Learning Techniques

As illustrated in Figure 5, defensive strategies such as adversarial training and robust learning techniques are critical in enhancing model resilience against adversarial attacks. The examples compare natural and adversarial training/testing accuracy across different perturbation budgets, address model overconfidence, and present deep learning processes for image generation and prediction. These examples emphasize defensive strategies' vital role in developing models that are both accurate and robust against adversarial challenges [73, 38, 74].

4.2 Dropout and Noise-Based Defenses

Dropout and noise-based defenses enhance AI model robustness against adversarial attacks by introducing stochasticity and variability into training and inference processes. Randomized smoothing and Gaussian noise integration within neural network layers aim to reduce the attack surface and bolster robustness. By embedding randomness into model architecture, these methods provide probabilistic resilience guarantees without direct exposure to adversarial examples [28, 16, 19].

Defensive dropout applies dropout during both training and testing phases, increasing gradient variance during adversarial example generation and complicating misclassification attempts [75]. Noise-based defenses further bolster robustness by introducing random noise into input data or model parameters, obscuring adversarial perturbation assessments and diminishing gradient-based attacks' effectiveness [38, 56, 16].

These strategies underscore variability and uncertainty's importance in AI systems as a defense mechanism, enhancing model resilience against adversarial attacks and contributing to robust AI systems capable of enduring diverse adversarial challenges. Findings indicate adversarial training improves model generalization and reduces prediction overconfidence, even on unperturbed data. Recent advancements emphasize generating natural and imperceptible adversarial examples, signaling a paradigm shift in adversarial research to address real-world security concerns and improve pretrained language model robustness [38, 28, 19]. As adversarial attack strategies evolve, refining and adapting dropout and noise-based defenses will be crucial for maintaining AI technologies' security and reliability.

4.3 GAN-Based Defense Strategies

Generative Adversarial Networks (GANs) are powerful tools in developing defense strategies, particularly in mitigating adversarial attack impacts. GANs generate synthetic data that enhances AI model robustness by providing diverse and challenging training samples. In presentation attack detection (PAD), GANs produce high-quality synthetic samples that improve training compared to traditional methods reliant solely on real data [76]. The AR-GAN framework reconstructs and classifies traffic sign images while mitigating adversarial attack effects, leveraging GANs' generative capabilities to counter adversarial perturbations [77].

GANs' application in defense strategies extends beyond image classification, demonstrating effectiveness in detecting and mitigating adversarial attacks across various domains. Recent methodologies leverage GAN's discriminator and generator to identify adversarial samples outside the learned data manifold, enabling a classifier-independent cleaning method. Approaches like PixelDefend utilize generative models to purify adversarially perturbed images, realigning them with the training data distribution, thus enhancing machine learning model resilience against diverse adversarial techniques [55, 78].

As adversarial attack techniques evolve, GAN-based defense strategies' significance, particularly those leveraging both GAN components to detect and mitigate adversarial samples, will grow increasingly critical for preserving AI systems' integrity and reliability. Recent research shows these strategies effectively identify adversarial inputs, consistently scoring them lower than legitimate data across various attacks and datasets, while enabling a cleaning method that projects adversarial samples back onto the data manifold [30, 78]. Ongoing refinement and application of GANs in security frameworks will be essential for safeguarding AI technologies against emerging threats, ensuring their robustness and trustworthiness across diverse applications.

4.4 Anomaly Detection and Prompt Engineering

Method Name	Defensive Techniques	Adversarial Threats	System Robustness
LM[79]	Anomaly Detection	Adversarial Examples	Withstand Adversarial Attacks
SRA[5]	Runtime Defenses	Backdoor Attacks	Maintaining Functionality
DD[10]	Adversarial Prompt Injections	Adversarial Examples	Maintain Performance
MHA[80]	Anomaly Detection	Data Poisoning	Withstand Adversarial Attacks
ADFAR[14]	Frequency-aware Randomization	Adversarial Word Substitution	Preserving Performance
AW[61]	Common Defense Mechanisms	Adversarial Examples	Higher Attack Success

Table 3: Summary of various defensive techniques and their effectiveness against different adversarial threats in AI systems. The table highlights methods such as anomaly detection and runtime defenses, detailing their capabilities to withstand attacks and maintain system robustness. References to specific studies provide context to the techniques and threats addressed.

Anomaly detection and prompt engineering are crucial in strengthening AI systems against adversarial threats, significantly enhancing defensive capabilities. Anomaly detection identifies deviations from expected behavior, indicating adversarial attacks or other security threats. Techniques like the Defensive Feature Layer (DFL) are instrumental in denoising perturbations within the feature space, reinforcing defenses against adversarial examples [81]. Training a surrogate Network Intrusion Detection System (NIDS) to identify optimal packet mutations that minimize the anomaly score exemplifies a defensive strategy through anomaly detection [79].

The importance of runtime defenses against deployment-stage backdoor attacks is underscored, suggesting prompt engineering could play a pivotal role in fortifying defenses [5]. Prompt engineering complements anomaly detection by strategically designing input prompts to mitigate adversarial threats. Inserting adversarial prompts into text to mislead large language models (LLMs) during inference illustrates prompt engineering's potential in enhancing defenses against unwanted data inference [10].

Moreover, integrating semantically relevant pixels, as demonstrated by methods like the Camouflager, enhances interpretability and can be incorporated into defensive strategies like adversarial training [80]. Frequency-Aware Randomization (ADFAR) combines randomization with anomaly detection to defend against adversarial word substitution, effectively mitigating threats without degrading non-adversarial sample performance [14]. Adversarial inputs can drive generative models to produce out-domain examples, highlighting vulnerabilities in GAN architectures [31]. Techniques like Adv-watermark create realistic adversarial examples carrying meaningful information, serving dual purposes of copyright protection and adversarial attack [61].

Integrating anomaly detection and prompt engineering into defense mechanisms significantly enhances AI systems' security and reliability, particularly in adversarial environments, by enabling malicious input identification and adaptation to evolving threats. This approach fortifies defenses against data poisoning and adversarial attacks while improving overall model robustness through

AI-generated, for reference only.

advanced algorithms that detect anomalous patterns and behaviors in real-time [46, 82, 28, 83]. As adversarial attack strategies continue to evolve, refining and integrating these techniques will be essential for maintaining AI technologies' integrity and safeguarding against emerging threats. Table 3 presents a comprehensive overview of defensive methods employed in AI systems to counteract adversarial threats, illustrating the effectiveness of various techniques in maintaining system robustness.

Feature	Adversarial Training and Robust Learning Techniques	Dropout and Noise-Based Defenses	GAN-Based Defense Strategies
Primary Mechanism	Adversarial Examples	Stochasticity	Synthetic Data
Defense Focus	Model Robustness	Gradient Variance	Sample Diversity
Unique Feature	Metric Learning	Randomized Smoothing	Pixeldefend

Table 4: Comparison of defensive strategies in AI systems, highlighting the primary mechanisms, defense focus, and unique features of adversarial training and robust learning techniques, dropout and noise-based defenses, and GAN-based defense strategies. This table provides a structured overview of the diverse approaches employed to enhance model resilience against adversarial attacks.

5 Prompt Engineering for Security

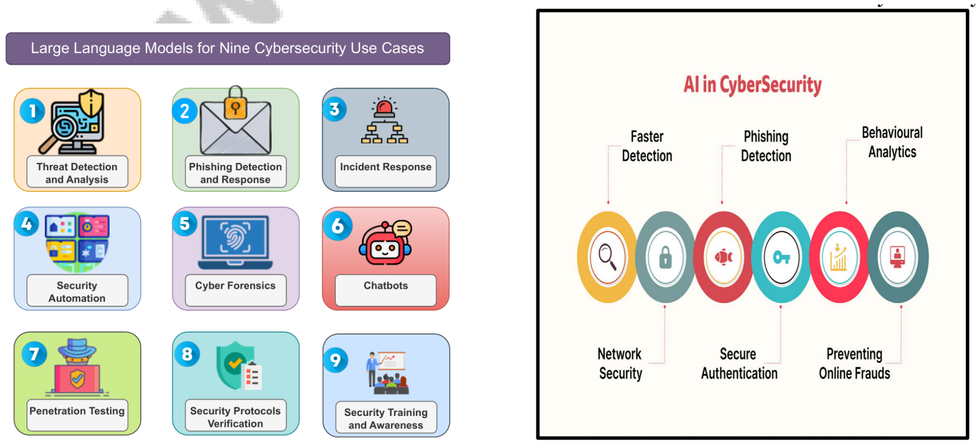
Prompt engineering is a critical strategy in advancing AI security measures, particularly in cybersecurity. This section explores its integration into various methodologies and frameworks, illustrating how prompt engineering enhances AI systems' resilience against adversarial threats. The discussion highlights its practical implications in addressing cybersecurity challenges.

5.1 Applications in Cybersecurity

Prompt engineering significantly bolsters AI systems' resilience against adversarial threats by strategically designing inputs. Cyber deception strategies, such as the Gaussian Mixture Model (GMM), introduce complexity and unpredictability, complicating potential attackers' efforts [84]. This proactive measure enhances AI systems' robustness within the cybersecurity landscape.

Adversarial sample generation techniques, like BMI-FGSM, reveal vulnerabilities in deep neural networks (DNNs) across applications, such as the Aliyun Image Recognition API, underscoring the need for prompt engineering in cybersecurity frameworks [85]. By integrating prompt engineering, cybersecurity measures can effectively anticipate and counteract adversarial strategies.

The Attacker Modeling Framework (AMF) simulates complex attacker behaviors, aiding in developing robust security measures that leverage prompt engineering principles [6]. AMF's integration exemplifies prompt engineering's critical role in anticipating and mitigating sophisticated adversarial threats.



(a) Large Language Models for Nine Cybersecurity Use Cases[86]

(b) AI in CyberSecurity[87]

Figure 6: Examples of Applications in Cybersecurity

AI-generated, for reference only.

As shown in Figure 6, prompt engineering, when integrated with AI and large language models (LLMs), is pivotal in cybersecurity. The first image illustrates nine distinct applications of LLMs in cybersecurity, showcasing their versatility in addressing complex security challenges. The second image highlights five key areas where AI enhances cybersecurity, emphasizing faster detection and behavioral analytics. Together, these examples underscore prompt engineering's transformative potential in fortifying cybersecurity frameworks [86, 87].

5.2 Evaluation and Effectiveness

Evaluating prompt engineering techniques for AI security requires integrating quantitative metrics and qualitative assessments. This dual approach is essential for measuring their efficacy against adversarial threats, such as prompt injection attacks in large language models (LLMs). Studies emphasize frameworks for assessing AI systems' resilience, including automated red teaming methods for identifying weaknesses [17, 83, 34, 33, 35]. Understanding how these techniques bolster model resilience is crucial.

Noise-aware training methods offer robustness against adversarial attacks without adversarial examples, achieving comparable results to traditional adversarial training [16]. Seatbelt-VAEs illustrate variational autoencoders' effectiveness in enhancing robustness while maintaining quality [88].

The tunable security framework demonstrates adaptability in AI security, reducing false positives while maintaining acceptable true positives [2]. FM-Defense achieves high detection accuracy against adversarial attacks, highlighting semantic understanding's importance in defense strategies [15].

Ensemble-based approaches effectively generate transferable targeted adversarial examples, indicating their potential in improving AI security against diverse strategies [12]. ADFAR outperforms existing methods against adversarial word substitution, maintaining accuracy for non-adversarial samples [14].

Ongoing evaluation and enhancement of prompt engineering techniques are vital for safeguarding AI systems' integrity against sophisticated threats like prompt injection attacks. Research highlights these attacks' potential to manipulate models, generating harmful content [89, 35]. Comprehensive understanding and robust testing frameworks are essential to counteract vulnerabilities, ensuring AI systems' intended functionality. By leveraging innovative defense strategies, AI security can be significantly enhanced, ensuring robust protection against sophisticated attacks.

6 Evaluation of AI Security

6.1 Innovative Frameworks and Models

Innovative frameworks and models are pivotal in evaluating AI security, offering structured methodologies to assess resilience against adversarial attacks. The Cyber Attack Thread model exemplifies this by providing a comprehensive framework for analyzing attack scenarios and corresponding security mechanisms, highlighting the dynamic nature of cyber threats and necessary countermeasures [49]. In adversarial machine learning, the Aux Block approach enhances neural network robustness against adversarial examples, enabling novel strategies to improve AI security [90]. Additionally, the benchmark by [91] facilitates comprehensive evaluations of attack detection mechanisms, exemplified by systems like Water Defense (WD).

Research into the resilience of generative models against adversarial attacks is crucial. The benchmark by [31] provides a framework for assessing generative model robustness, essential for understanding vulnerabilities and developing mitigation strategies. Experiments utilizing ImageNet and CASIA-WebFace datasets showcase Adv-watermark's performance against baseline black-box attack methods [61], emphasizing the need for robust evaluation models.

Future research may focus on optimizing proposal distributions for specific models and enhancing decision-based attack efficiency [56], which is vital for advancing AI security. Evaluations involving 255,000 power quality signals across 17 classes underscore the importance of comprehensive testing environments in understanding and mitigating adversarial threats [3]. Collectively, these frameworks and models significantly enhance our understanding of AI security, advocating for collaborative efforts to navigate the complexities of this domain [17, 83, 19]. As AI security evolves, refining these frameworks will be crucial for safeguarding technologies against emerging threats.

AI-generated, for reference only.

6.2 Evaluation Metrics for Adversarial Attacks

Evaluating adversarial attacks on AI systems necessitates a comprehensive set of metrics. The Attack Success Rate (ASR) measures the proportion of adversarial examples that successfully induce misclassification, while the Semantics Preservation Rate (SPR) ensures adversarial examples maintain their intended semantic meaning [53]. Robust accuracy evaluates defense mechanisms' ability to maintain accuracy amidst adversarial perturbations, focusing on reducing false positives and ensuring computational efficiency [2]. Adversarial accuracy, assessed under various attacks like PGD and FGSM, is often measured using Area Under Curve (AUC) scores, providing a nuanced understanding of model robustness [45].

For large language models, metrics such as Clean Accuracy (CA), Trigger Accuracy (TA), Triggered Accuracy Ratio (TAR), Average Shannon Entropy (ASE), and Randomized Attack Success Rate (RASR) evaluate architectural backdoor attacks' effectiveness [13]. These metrics offer a comprehensive view of the attack's impact on model performance and security. Transferability of adversarial examples complicates defense efforts, necessitating robust strategies to measure their success in misleading black-box models [12]. Detection accuracy, gauged by the proportion of correctly identified adversarial and clean instances, is pivotal for evaluating defense strategies, particularly the effectiveness of purification methods [15].

Evaluating data defenses involves measuring attackers' failure rate to infer personal information, achieving over 90

6.3 Frameworks and Benchmarks for Security Evaluation

Assessing AI security relies on diverse frameworks and benchmarks that provide structured methodologies for evaluating robustness against adversarial attacks. The Cyber Attack Thread model offers a comprehensive structure for analyzing attack scenarios and required security measures [49]. In adversarial machine learning, the Aux Block approach enhances model robustness against adversarial examples by embedding additional defensive layers within the architecture [90]. The benchmark introduced by [91] allows comprehensive assessments of attack detection mechanisms, critical for understanding defense strategy effectiveness.

Research on generative models' resilience to adversarial attacks is essential. The benchmark by [31] provides insights into assessing generative model robustness and identifying vulnerabilities. Experiments with datasets like ImageNet and CASIA-WebFace highlight Adv-watermark's performance against baseline black-box attack methods [61], emphasizing the need for robust evaluation models. Future research could optimize proposal distributions for specific models and enhance decision-based attack efficiency [56], vital for advancing AI security. Evaluations involving 255,000 power quality signals across 17 classes illustrate the effectiveness of proposed adversarial methods in assessing AI security [3].

Integrating various frameworks and benchmarks enhances comprehension of AI security, providing innovative methodologies for assessing and strengthening resilience against adversarial threats. This collective effort addresses practical concerns in multiple domains and emphasizes real-world applicability, as seen in initiatives like the Advbench dataset and unified frameworks for constrained adversarial attacks [17, 92, 47, 19]. As AI security evolves, refining these frameworks will be critical for safeguarding technologies against emerging threats.

6.4 Methodologies for Defensive Strategy Assessment

Assessing defensive strategies in AI security involves evaluating the effectiveness of various mechanisms against adversarial threats. Robust methodologies are essential to ensure resilience to evolving attack vectors. Empirical evaluations of classifiers' security under simulated attack scenarios provide insights into vulnerabilities and defense effectiveness [22]. This empirical approach is complemented by methodologies that consider all potential attack vectors in risk evaluations for comprehensive assessments. The Cowboy methodology emphasizes a generalizable approach to evaluating defensive strategies across different attack methods and datasets, highlighting simplicity and adaptability [78]. This method underscores the importance of tailored risk assessments based on the specific context of the AI system evaluated.

In language models, performance evaluation compares original accuracy on non-adversarial samples to accuracy after adversarial attacks, essential for understanding adversarial impact [14]. Critical-time metrics formulated within a quadratic constraints framework enhance the evaluation of protection, detection, and reaction times available for defenders against adversarial attacks [4, 93]. Analyzing the timing of security incidents informs when defensive actions should be initiated to effectively counteract breaches.

The Sludge Strategy evaluates qualitative and quantitative costs imposed on attackers, highlighting the operational effects on adversaries when formulating defensive strategies, particularly in light of reconnaissance techniques employed throughout the cyber kill chain [94, 93]. Methodologies for assessing defensive strategies are diverse and continuously evolving, driven by the complexity of adversarial threats. This evolution is underscored by the sophistication of cyber attacks that exploit vulnerabilities identified through adversarial reconnaissance techniques, necessitating AI integration for enhanced threat detection and response capabilities [17, 19, 29, 93, 87]. Comprehensive assessment techniques ensure the robustness and reliability of AI systems against emerging security challenges.

6.5 Case Studies and Comparative Analyses

Case studies and comparative analyses are crucial for understanding the efficacy of adversarial attack and defense strategies across AI models and datasets. The TextAttack framework illustrates the importance of structured evaluation frameworks in identifying vulnerabilities and improving defenses against adversarial threats in state-of-the-art natural language processing (NLP) models [47]. The ZK-GanDef framework provides a comparative analysis of zero knowledge defenses, achieving significant test accuracy improvements on adversarial examples by up to 49.17

Experiments on vision-language models (VLMs) reveal their susceptibility to adversarial pop-ups, underscoring the need for continuous research into robust defense mechanisms [95]. A systematic survey by Navarro et al. offers a comparative analysis of detection methods, focusing on their effectiveness in identifying multi-step attacks and implementation challenges [96]. This analysis guides the development of more effective defense mechanisms. The survey by Sangwan et al. compares various AI attacks and defense mechanisms, highlighting their effectiveness and limitations in mitigating vulnerabilities [23]. In cyber-physical systems, case studies on sensor deception attacks demonstrate the effectiveness of the Automated Adversary Emulation (AAE) method, offering a basis for comparison with existing evaluation methods [97]. Experiments using benchmark datasets like MNIST, CIFAR-10, and CelebA assess hijacked models against clean models, providing insights into utility and attack success rates [80].

The proposed CLAP method showcases significant improvements in generating behavior-diverse agents for penetration testing, outperforming state-of-the-art methods in various environments [98]. The benchmark presented by Alam et al. demonstrates accuracy in extracting relevant attack patterns, serving as a valuable tool for proactive defense against emerging threats [25]. Future work aims to refine probabilistic models and break down actions into smaller components for improved planning, enhancing adversarial strategies [99]. Additionally, the proposed statistical detector shows superior performance in identifying adversarial examples, highlighting its compatibility with other defenses [46].

These case studies and comparative analyses provide critical insights into the adversarial landscape, enhancing understanding of various evaluation methods and outcomes. Ongoing assessment and enhancement of defense strategies enable researchers to create more robust AI systems against sophisticated adversarial threats. This iterative process involves leveraging AI for threat detection, vulnerability assessment, and incident response, ensuring effective risk mitigation. Understanding the ethical and technical challenges associated with AI in cybersecurity is essential for developing responsible solutions, ultimately ensuring AI-driven defenses can adapt to the evolving cyber threat landscape [17, 18, 28, 19, 29].

7 Conclusion

7.1 Challenges in AI Security

The landscape of AI security is fraught with significant challenges, primarily due to the rapid evolution of adversarial threats and the limitations inherent in current defense mechanisms. A major obstacle is the development of defense methods that can generalize across diverse adversarial attacks, as the robustness of AI models is highly contingent on the characteristics of their training data. This variability complicates the creation of universal defense strategies capable of countering all potential threats effectively. The persistent threat of stealthy backdoors in AI models further complicates security efforts, as these vulnerabilities can remain undetected and resilient against traditional defenses. The ability of attackers to embed such undetectable backdoors underscores the pressing need for advanced detection techniques to effectively identify and neutralize these threats.

Future research should prioritize enhancing the stealthiness of adversarial attacks, such as those driven by explanation-based methods, and focus on developing defenses against them. Accurately modeling attacker behavior adds another layer of complexity to formulating effective security measures, necessitating comprehensive frameworks that consider the diverse strategies employed by adversaries. The emergence of deployment-stage backdoor attacks, including those exploiting vulnerabilities during the deployment process, highlights the necessity for robust defenses to maintain the integrity of deep neural networks (DNNs). These challenges emphasize the importance of continuous evaluation and refinement of security measures.

Despite advancements in adversarial robustness and detection efficiency, achieving optimal performance under adversarial conditions remains a formidable challenge. Practical challenges in implementing effective machine learning security measures are often overlooked, leading to gaps in understanding real-world vulnerabilities and the applicability of theoretical solutions. Proposed data defenses offer a cost-effective means for users to protect against unwanted inference, but they do not guarantee absolute security. This underscores the need for ongoing research to develop more comprehensive security measures that can adapt to the evolving landscape of adversarial threats.

Addressing these challenges requires a multifaceted approach, including the development of adaptive defenses, comprehensive evaluation frameworks, and robust training methodologies to protect AI systems against a variety of adversarial attacks. By enhancing the adaptability and efficiency of defensive strategies, AI systems can be better equipped to withstand the complexities of evolving adversarial landscapes.

7.2 Challenges and Future Directions in Security Evaluation

The rapidly changing landscape of AI security evaluation presents numerous challenges, particularly in adapting to the sophisticated nature of adversarial threats. A significant challenge is the creation of adaptive defenses capable of classifying and responding to various types of noise, thereby improving detection strategies for assessing attack strength. This adaptability is crucial for enhancing AI systems' resilience against a broad spectrum of adversarial attacks. Future research should focus on refining machine learning models to effectively withstand adversarial assaults, emphasizing the importance of data diversity in training processes to enhance model robustness.

Moreover, the intersection of AI and cybersecurity introduces unique challenges, necessitating the development of robust security protocols to address new attack vectors in AI applications. Establishing comprehensive taxonomies of reconnaissance techniques is essential for bolstering defensive strategies in AI security, providing a structured framework for understanding and mitigating adversarial reconnaissance activities.

Empirical validation of existing benchmarks and their expansion to cover a wider array of machine learning applications and attack scenarios is necessary to ensure the relevance and applicability of security evaluations. Developing adaptable defense mechanisms and improving the understanding of adversarial strategies remain critical areas for future research. Enhancing detection accuracy and creating adaptive models to tackle the complexities of multi-step attacks are vital for advancing AI security evaluations. Additionally, addressing ethical considerations in AI deployment, especially in cybersecurity, is crucial for balancing technological advancement with ethical guidelines.

Future research directions should also include extending the critical-time method to address stealthy attacks and integrating specific communication models, highlighting ongoing challenges in AI security evaluation. Investigating the psychological traits of attackers and exploring different types of sludge can further enhance cybersecurity measures. Expanding benchmarks to incorporate adaptive attack scenarios and refining defense techniques are promising areas for future exploration. By addressing these challenges and pursuing these research directions, the field of AI security evaluation can progress toward more robust and reliable systems capable of withstanding sophisticated adversarial threats.

References

[1] Yingdi Wang, Wenjia Niu, Tong Chen, Yingxiao Xiang, Jingjing Liu, Gang Li, and Jiqiang Liu. A training-based identification approach to vin adversarial examples, 2018.

[2] Omer Katz and Benjamin Livshits. Security: Doing whatever is needed... and not a thing more!, 2018.

[3] Jiwei Tian, Buhong Wang, Jing Li, Zhen Wang, and Mete Ozay. Adversarial attacks and defense methods for power quality recognition, 2022.

[4] Sadegh Farhang and Jens Grossklags. When to invest in security? empirical evidence and a game-theoretic approach for time-based security, 2017.

[5] Xiangyu Qi, Jifeng Zhu, Chulin Xie, and Yong Yang. Subnet replacement: Deployment-stage backdoor attack against deep neural networks in gray-box setting, 2021.

[6] Christopher Deloglos, Carl Elks, and Ashraf Tantawy. An attacker modeling framework for the assessment of cyber-physical systems security, 2021.

[7] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Li Chen, Michael E. Kounavis, and Duen Horng Chau. Adagio: Interactive experimentation with adversarial attack and defense for audio, 2018.

[8] Kathrin Grosse, Lukas Bieringer, Tarek Richard Besold, Battista Biggio, and Katharina Krombholz. Machine learning security in industry: A quantitative survey, 2023.

[9] Linan Huang and Quanyan Zhu. Strategic learning for active, adaptive, and autonomous cyber defense, 2019.

[10] William Agnew, Harry H. Jiang, Cella Sum, Maarten Sap, and Sauvik Das. Data defenses against large language models, 2024.

[11] Derui Wang, Chaoran Li, Sheng Wen, Qing-Long Han, Surya Nepal, Xiangyu Zhang, and Yang Xiang. Daedalus: Breaking non-maximum suppression in object detection via adversarial examples, 2020.

[12] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks, 2017.

[13] Abdullah Arafat Miah and Yu Bi. Exploiting the vulnerability of large language models via defense-aware architectural backdoor, 2024.

[14] Rongzhou Bao, Jiayi Wang, and Hai Zhao. Defending pre-trained language models from adversarial word substitutions without performance sacrifice, 2021.

[15] Shuo Wang, Tianle Chen, Surya Nepal, Carsten Rudolph, Marthie Grobler, and Shangyu Chen. Defending adversarial attacks via semantic feature manipulation, 2020.

[16] Ayoub Arous, Andres F Lopez-Lopera, Nael Abu-Ghazaleh, and Ihsen Alouani. May the noise be with you: Adversarial training without adversarial examples, 2023.

[17] Fnu Jimmy. Emerging threats: The latest cybersecurity risks and the role of artificial intelligence in enhancing cybersecurity defenses. *Valley International Journal Digital Library*, 1:564–74, 2021.

[18] Nicolas Guzman Camacho. The role of ai in cybersecurity: Addressing threats in the digital age. *Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023*, 3(1):143–154, 2024.

[19] Yangyi Chen, Hongcheng Gao, Ganqu Cui, Fanchao Qi, Longtao Huang, Zhiyuan Liu, and Maosong Sun. Why should adversarial perturbations be imperceptible? rethink the research paradigm in adversarial nlp, 2022.

[20] Daryna Oliynyk, Rudolf Mayer, and Andreas Rauber. I know what you trained last summer: A survey on stealing machine learning models and defences, 2023.

[21] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Chau. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*, pages 52–68. Springer, 2019.

[22] Battista Biggio, Giorgio Fumera, and Fabio Roli. Security evaluation of pattern classifiers under attack, 2017.

[23] Raghvinder S Sangwan, Youakim Badr, and Satish M Srinivasan. Cybersecurity for ai systems: A survey. *Journal of Cybersecurity and Privacy*, 3(2):166–190, 2023.

[24] Chris Lu, Timon Willi, Alistair Letcher, and Jakob Foerster. Adversarial cheap talk, 2023.

[25] Md Tanvirul Alam, Dipkamal Bhusal, Youngja Park, and Nidhi Rastogi. Looking beyond iocs: Automatically extracting attack patterns from external cti, 2023.

[26] Kun Li, Fan Zhang, and Wei Guo. Atwm: Defense against adversarial malware based on adversarial training, 2023.

[27] Kiana Vu, Phung Lai, and Truc Nguyen. Xsub: Explanation-driven adversarial attack against blackbox classifiers via feature substitution, 2024.

[28] Jiayi Wang, Rongzhou Bao, Zhuosheng Zhang, and Hai Zhao. Rethinking textual adversarial defense for pre-trained language models, 2022.

[29] Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. Bert-attack: Adversarial attack against bert using bert. *arXiv preprint arXiv:2004.09984*, 2020.

[30] Tao Yu, Shengyuan Hu, Chuan Guo, Wei-Lun Chao, and Kilian Q. Weinberger. A new defense against adversarial images: Turning a weakness into a strength, 2019.

[31] Dario Pasquini, Marco Mingione, and Massimo Bernaschi. Adversarial out-domain examples for generative models, 2019.

[32] Qing Song, Yingqi Wu, and Lu Yang. Attacks on state-of-the-art face recognition using attentional adversarial attack generative network, 2018.

[33] Wenxiao Zhang, Xiangrui Kong, Conan Dewitt, Thomas Braunl, and Jin B. Hong. A study on prompt injection attack against llm-integrated mobile robotic systems, 2024.

[34] Bojian Jiang, Yi Jing, Tianhao Shen, Tong Wu, Qing Yang, and Deyi Xiong. Automated progressive red teaming, 2024.

[35] Xiaogeng Liu, Zhiyuan Yu, Yizhe Zhang, Ning Zhang, and Chaowei Xiao. Automatic and universal prompt injection attacks against large language models, 2024.

[36] Ying Yuan, Giovanni Apruzzese, and Mauro Conti. Multi-spacephish: Extending the evasion-space of adversarial attacks against phishing website detectors using machine learning, 2023.

[37] Marius Mosbach, Maksym Andriushchenko, Thomas Trost, Matthias Hein, and Dietrich Klakow. Logit pairing methods can fool gradient-based attacks, 2019.

[38] Robust models are less over-confident.

[39] Ayush Goel. An empirical review of adversarial defenses, 2020.

[40] Xinlei He, Zheng Li, Weilin Xu, Cory Cornelius, and Yang Zhang. Membership-doctor: Comprehensive assessment of membership inference against machine learning models, 2022.

[41] Yifan Shen, Zhengyuan Li, and Gang Wang. Practical region-level attack against segment anything models, 2024.

[42] Marco Alecci, Mauro Conti, Francesco Marchiori, Luca Martinelli, and Luca Pajola. Your attack is too dumb: Formalizing attacker scenarios for adversarial transferability, 2023.

[43] Akshita Jha and Chandan K. Reddy. Codeattack: Code-based adversarial attacks for pre-trained programming language models, 2023.

[44] Fuxun Yu, Zhuwei Qin, Chenchen Liu, Liang Zhao, Yanzhi Wang, and Xiang Chen. Interpreting and evaluating neural network robustness, 2019.

[45] Chengzhi Mao, Ziyuan Zhong, Junfeng Yang, Carl Vondrick, and Baishakhi Ray. Metric learning for adversarial robustness, 2019.

[46] Alessandro Cennamo, Ido Freeman, and Anton Kummert. A statistical defense approach for detecting adversarial examples, 2019.

[47] John X. Morris, Jin Yong Yoo, and Yanjun Qi. Textattack: Lessons learned in designing python frameworks for nlp, 2020.

[48] Arthur Perodou, Christophe Combastel, and Ali Zolghadri. Critical-time metric for risk analysis against sharp input anomalies: computation and application case study, 2023.

[49] Koustav Sadhukhan, Rao Arvind Mallari, and Tarun Yadav. Cyber attack thread: A control-flow based approach to deconstruct and mitigate cyber threats, 2016.

[50] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan Yuille. Improving transferability of adversarial examples with input diversity, 2019.

[51] Han Qiu, Yi Zeng, Qinkai Zheng, Tianwei Zhang, Meikang Qiu, and Gerard Memmi. Mitigating advanced adversarial attacks with more advanced gradient obfuscation techniques, 2020.

[52] Gilad Cohen, Guillermo Sapiro, and Raja Giryes. Detecting adversarial samples using influence functions and nearest neighbors, 2020.

[53] Xiang Ling, Zhiyu Wu, Bin Wang, Wei Deng, Jingzheng Wu, Shouling Ji, Tianyue Luo, and Yanjun Wu. A wolf in sheep's clothing: Practical black-box adversarial attacks for evading learning-based windows malware detection in the wild, 2024.

[54] Hui Liu, Bo Zhao, Minzhi Ji, and Peng Liu. Greedyfool: Multi-factor imperceptibility and its application to designing a black-box adversarial attack, 2021.

[55] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples, 2018.

[56] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*, 2017.

[57] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.

[58] Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. Gradient-based adversarial attacks against text transformers, 2021.

[59] Josiah Dykstra, Kelly Shortridge, Jamie Met, and Douglas Hough. Sludge for good: Slowing and imposing costs on cyber attackers, 2022.

[60] Fnu Suya, Jianfeng Chi, David Evans, and Yuan Tian. Hybrid batch attacks: Finding black-box adversarial examples with limited queries, 2019.

[61] Xiaojun Jia, Xingxing Wei, Xiaochun Cao, and Xiaoguang Han. Adv-watermark: A novel watermark perturbation for adversarial examples, 2020.

[62] Yulong Wang, Tianxiang Li, Shenghong Li, Xin Yuan, and Wei Ni. New adversarial image detection based on sentiment analysis, 2023.

[63] Sayantan Sarkar, Ankan Bansal, Upal Mahbub, and Rama Chellappa. Upset and angri : Breaking high performance image classifiers, 2017.

[64] Fanyou Wu, Rado Gazo, Eva Haviarova, and Bedrich Benes. Efficient project gradient descent for ensemble adversarial attack, 2019.

[65] Bin Chen, Jia-Li Yin, Shukai Chen, Bo-Hao Chen, and Ximeng Liu. An adaptive model ensemble adversarial attack for boosting adversarial transferability, 2023.

[66] Yifeng Xiong, Jiadong Lin, Min Zhang, John E. Hopcroft, and Kun He. Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability, 2022.

[67] Ruoxi Qin, Linyuan Wang, Xingyuan Chen, Xuehui Du, and Bin Yan. Dynamic defense approach for adversarial robustness in deep neural networks via stochastic ensemble smoothed model, 2021.

[68] Gunnar Mein, Kevin Hartman, and Andrew Morris. Firebert: Hardening bert-based classifiers against adversarial attack, 2020.

[69] Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. Word-level textual adversarial attacking as combinatorial optimization, 2020.

[70] Kaizhao Liang, Jacky Y. Zhang, Boxin Wang, Zhuolin Yang, Oluwasanmi Koyejo, and Bo Li. Uncovering the connections between adversarial transferability and knowledge transferability, 2021.

[71] Ehsan Nowroozi, Yassine Mekdad, Mohammad Hajian Berenjestanaki, Mauro Conti, and Abdeslam EL Fergougui. Demystifying the transferability of adversarial attacks in computer networks, 2022.

[72] Zelin Li, Kehai Chen, Lemao Liu, Xuefeng Bai, Mingming Yang, Yang Xiang, and Min Zhang. Tf-attack: Transferable and fast adversarial attacks on large language models, 2024.

[73] Chaojian Yu, Dawei Zhou, Li Shen, Jun Yu, Bo Han, Mingming Gong, Nannan Wang, and Tongliang Liu. Strength-adaptive adversarial training, 2022.

[74] Haojing Shen, Sihong Chen, Ran Wang, and Xizhao Wang. Adversarial learning with cost-sensitive classes, 2022.

[75] Siyue Wang, Xiao Wang, Pu Zhao, Wujie Wen, David Kaeli, Peter Chin, and Xue Lin. Defensive dropout for hardening deep neural networks under adversarial attacks, 2018.

[76] Reuben Markham, Juan M. Espin, Mario Nieto-Hidalgo, and Juan E. Tapia. Open-set: Id card presentation attack detection using neural transfer style, 2023.

[77] M Sabbir Salek, Abdullah Al Mamun, and Mashrur Chowdhury. Ar-gan: Generative adversarial network-based defense method against adversarial attacks on the traffic sign classification system of autonomous vehicles, 2023.

[78] Gokula Krishnan Santhanam and Paulina Grnarova. Defending against adversarial attacks by leveraging an entire gan, 2018.

[79] Ke He, Dan Dongseong Kim, Jing Sun, Jeong Do Yoo, Young Hun Lee, and Huy Kang Kim. Liuer mihou: A practical framework for generating and evaluating grey-box adversarial attacks against nids, 2022.

[80] Ahmed Salem, Michael Backes, and Yang Zhang. Get a model! model hijacking attack against machine learning models, 2021.

[81] Mohammed Hassanin, Ibrahim Radwan, Nour Moustafa, Murat Tahtali, and Neeraj Kumar. Mitigating the impact of adversarial attacks in very deep networks, 2020.

[82] Andrea Paudice, Luis Muñoz-González, Andras Gyorgy, and Emil C. Lupu. Detection of adversarial training examples in poisoning attacks through anomaly detection, 2018.

[83] Sakthiswaran Rangaraju. Ai sentry: Reinventing cybersecurity through intelligent threat detection. *EPH-International Journal of Science And Engineering*, 9(3):30–35, 2023.

[84] Linan Huang and Quanyan Zhu. Duplicity games for deception design with an application to insider threat mitigation, 2021.

[85] Junyu Lin, Lei Xu, Yingqi Liu, and Xiangyu Zhang. Black-box adversarial sample generation based on differential evolution, 2020.

[86] Mohamed Amine Ferrag, Fatima Alwahedi, Ammar Battah, Bilel Cherif, Abdechakour Mechri, Norbert Tihanyi, Tamas Bisztray, and Merouane Debbah. Generative ai in cybersecurity: A comprehensive review of llm applications and vulnerabilities, 2025.

[87] Asad Yaseen. Ai-driven threat detection and response: A paradigm shift in cybersecurity. *International Journal of Information and Cybersecurity*, 7(12):25–43, 2023.

[88] Matthew Willetts, Alexander Camuto, Tom Rainforth, Stephen Roberts, and Chris Holmes. Improving vaes' robustness to adversarial attack, 2021.

[89] Zhongjie Ba, Jieming Zhong, Jiachen Lei, Peng Cheng, Qinglong Wang, Zhan Qin, Zhibo Wang, and Kui Ren. Surrogateprompt: Bypassing the safety filter of text-to-image models via substitution, 2024.

[90] Yueyao Yu, Pengfei Yu, and Wenye Li. Auxblocks: Defense adversarial example via auxiliary blocks, 2019.

[91] Sridhar Adepu and Aditya Mathur. Assessing the effectiveness of attack detection at a hackfest on industrial control systems, 2018.

[92] Thibault Simonetto, Salijona Dyrmishi, Salah Ghamizi, Maxime Cordy, and Yves Le Traon. A unified framework for adversarial attack and defense in constrained feature space, 2022.

[93] Shanto Roy, Nazia Sharmin, Jaime C. Acosta, Christopher Kiekintveld, and Aron Laszka. Survey and taxonomy of adversarial reconnaissance techniques, 2022.

[94] Ron Bitton, Nadav Maman, Inderjeet Singh, Satoru Momiyama, Yuval Elovici, and Asaf Shabtai. Evaluating the cybersecurity risk of real world, machine learning production systems, 2021.

[95] Yanzhe Zhang, Tao Yu, and Diyi Yang. Attacking vision-language computer agents via pop-ups, 2024.

[96] Julio Navarro, Aline Deruyver, and Pierre Parrend. A systematic survey on multi-step attack detection. *Computers & Security*, 76:214–249, 2018.

[97] Arnab Bhattacharya, Thiagarajan Ramachandran, Sandeep Banik, Chase P. Dowling, and Shaunak D. Bopardikar. Automated adversary emulation for cyber-physical systems via reinforcement learning, 2020.

[98] Yizhou Yang and Xin Liu. Behaviour-diverse automatic penetration testing: A curiosity-driven multi-objective deep reinforcement learning approach, 2022.

[99] Carlos Sarraute, Gerardo Richarte, and Jorge Lucangeli Obes. An algorithm to find optimal attack paths in nondeterministic scenarios, 2013.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.