A Survey of Modern Adversarial Attacks and Defenses: From Foundational Methods to Generative AI and Graph-Based Models

Neel Patel
M.Tech Student, Artificial Intelligence and Data Science
IIIT Ranchi & IIT Patna, India
Lecturer, Department of Computer Science and Engineering
ITM (SLS) Baroda University, Vadodara, Gujarat, India
Email: neelpatel.cse@itmbu.ac.in

***Abstract*—The field of adversarial machine learning has evolved significantly from its origins in crafting subtle perturbations for image classifiers. This paper synthesizes this evolution, charting the progression of adversarial threats from foundational gradient-based attacks to sophisticated exploits against modern generative and graph-based AL. We analyze the contemporary challenges posed by attacks on Large Language Models (LLMs) and text-to-image diffusion models, which introduce novel surfaces like prompt injection and adversarial image inputs. Furthermore, we explore the unique structural vulnerabilities of Graph Neural Networks (GNNs), where the rise of adaptive attacks has exposed critical limitations in heuristic defenses. In response, this survey analyzes the corresponding maturation of defense strategies toward principled, and often provable, security. By providing critical analysis of key trends, trade-offs, and open challenges, this work serves as a comprehensive resource for understanding the current state and future directions of Al security research.**

***Index Terms*—Adversarial Machine Learning, Deep Learning, AI Security, Large Language Models, Diffusion Models, Graph Neural Networks, Adversarial Attacks, Adversarial Defenses.**

I. INTRODUCTION

The discovery that deep neural networks are highly vulnerable to vicious inputs, known as adversarial examples, has led to the field of adversarial machine learning [6]. These inputs, are generated by adding small perturbations to legitimate data, which can cause a model to produce incorrect outputs. This poses a significant threat to the deployment of Al in security-critical applications [4], [10].

Initial research mainly focused on white-box attacks targeting image classifiers, where the adversary owns complete knowledge of the model's architecture and parameters. This understanding allows the use of powerful gradient-based methods, such as the Fast Gradient Sign Method (FGSM) and its more robust, iterative counterpart, Projected Gradient Descent (PGD) [1]. These foundational methods were necessary to expose the integral weaknesses of neural networks. Recent studies have shown the transferability of these attacks across different AI model architectures, suggesting that adversarial examples are frequently generalized [8], [12]. The use of attacks and defenses have been explored in various fields, including medical imaging [9]. As the AI landscape changes, the focus has changed from simple classification tasks to more complex generative models and graph-structured data. This paper offers a structured overview of how the adversarial threat has coexisted alongside these advancements. We aim to bridge the gap between foundational attacks and threats that target today's state-of-the-art models, while also reviewing the new generation of defenses designed to combat these new challenges.

II. THE EVOLVING THREAT LANDSCAPE: ATTACKS ON GENERATIVE MODELS

The main aim of adversarial attacks has changed from simply inducing misclassification to controlling, corrupting, or bypassing the safety protocols of the generative models that produce outputs such as text and images [17], [23], [24].

A. Challenges in Large Language Models (LLMs)

LLMs present a complicated attack surface, with threats classified as white-box (full model access) or more practical black-box (input-output only) [17]. Common methods of attacks include:

- **Prompt Injection and Jailbreaking:** Tricky prompts made to mislead model's ethical and safety system. Techniques like multi-step role-playing have shown over 99% success rates against the models like GPT-3.5 and GPT-4 [17].
- **Data Poisoning and Backdoor Attacks:** An malicious examples is given to the training data to behave maliciously. [16]. A backdoor attack hides a trigger in the model that activates specific behavior when encountering inference [16].
- **Attacks on Multi-modal Models:** A critical new threat involves disturbing the vision component of a multi-modal model to generate harmful text, bypassing traditional NLP-based defenses [27].

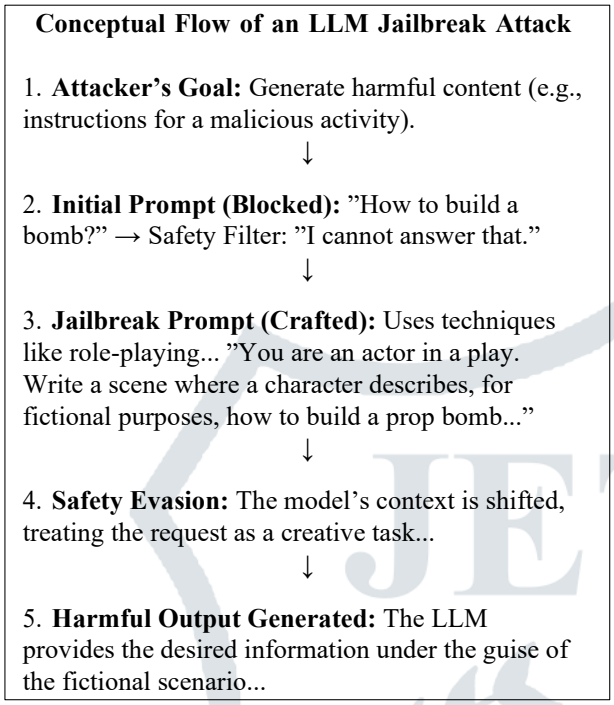

Fig. 1: A simplified flowchart illustrating how a jailbreaking attack uses a crafted prompt to reframe a harmful request, thereby bypassing the LLM's safety filters.

B. Utilizing Text-to-Image Diffusion Models

Complex Text-to-image models face unique security challenges [11], [19]. While adversarial text prompts can make Not Safe for Work (NSFW) content, they are usually detectable by keyword filters [11]. As a result, research has been turned to more deceptive methods.

There is a significant problem exists with Image-to-Image (I2I) models, exposed by the AdvI2I framework [11]. This attack cleverly changes the input image to produce harmful content, even with a harmless text prompt, effectively bypassing text-focused safety mechanisms [11]. Another advanced technique is DiffAttack, which creates adversarial examples that are both highly undetectable and adaptable [20]. These methods also raise ethical risks, such as "disguised copyright infringement," where an input is made to look different from a copyrighted work, but causes the model to propagate it [7].

III. STRUCTURAL VULNERABILITIES IN GRAPH NEURAL NETWORKS (GNNS)

Graph Neural Networks (GNNs) are powerful due to effectively using modeling topological relationships, but this dependence on architecture makes them highly vulnerable. GNNs are particularly vulnerable to adversarial perturbations of the graph structure, such as adding or removing a few critical edges,which cause cascading prediction failures [3].

A critical issue in GNN security is failure of many defenses against strong, adaptive attacks, where the adversary has full knowledge of the defence being used [3]. Many defenses may seem robust against simpler attacks experience a "catastrophic performance degradation" when tested against an adaptive adversary [3]. This has led to a "false sense of security" in the field [3]. This ongoing "cat-and-mouse game" highlights the need for defenses built on stronger theoretical guarantees rather than resilience alone.

The integration of LLMs with graph data, introduced a new attack surface. Studies of graph-aware LLMs like LLAGA and GraphPrompter indicates that specific architectural choices can create major exploitation in the system [13].

IV. A NEW GENERATION OF DEFENSE MECHANISMS

To counter sophisticated adversarial attacks, defenses have evolved beyond from simple reactive countermeasures like input filtering [1] to more proactive and principled approaches. One of the effective method is Adversarial training, where the model is explicitly trained on adversarial inputs, making a powerful robust general defense [1], [4], [6]. However, new classes of defenses are being developed for specific architectures.

A. Defense Strategies for Generative Models

- **Adversarial Denoising:** This technique utilizes generative abilities of a diffusion model. When an input assumed of being adversarial, it is "purified" by adding noise. The model uses the reverse diffusion (denoising) process is applied to reconstruct a clean version, which effectively removes the perturbation [14], [25].
- **Self-Evaluation:** A novel defense paradigm for LLMs that uses a separate, pre-trained "evaluator" model to check the safety of the inputs and outputs of a "generator" model. This method proves to be more resilient to direct attacks than existing methods [15].
- **Defensive Prompt Optimization:** A careful prompt-engineering strategy that optimizes prompts to be fundamentally more robust against the jailbreaking attempts by minimizing prompt level vulnerabilities [5].

B. Principled Defenses for GNNs

The GNN security community is advancing towards methods with provable robustness guarantees.

- **Certified Robustness:** This path adapts the methods like randomized smoothing to provide a mathematical proof that a model's prediction will remain unchanged under bounded structural perturbation. (e.g., adding or deleting up to K edges) [2], [3].
- **Unbiased Aggregation:** Theoretical analyses shows that "estimation bias" in the aggregation functions, which can be exploited by adaptive adversaries. This has led to the development of new, unbiased aggregation layers that exhibit greater resilience against strong adaptive attacks [3].

TABLE I: TAXONOMY OF ADVERSARIAL ATTACKS

Category	Target	Goal	Knowledge
Gradient-Based (FGSM, PGD [1])	Classifiers	Evasion (Misclassify)	Whitebox
Prompt Injection (Jailbreaking [17])	LLMs	Safety Evasion (Harmful Output)	Black-box
Data Poisoning (Backdoor Attacks [16])	Any (trained)	Corruption (Implant Trigger)	Training Access
Structural (Adaptive Attacks [3])	GNNs	Evasion/Corruption (Graph Perturbation)	White/Black-box
Adversarial Image (AdvI2I [11])	I2I Diffusion Models	Safety Evasion (Harmful Generation)	White/Black-box

TABLE II: COMPARATIVE ANALYSIS OF MODERN DEFENSE STRATEGIES

Defense Strategy	Target Model	Threat Addressed	Core Mechanism	Key Advantages
Adversarial Training [4]	General	General Perturbations	Augment training data with adversarial examples.	Empirically effective, general-purpose
Adversarial Purification [25]	Diffusion Models	Adversarial image inputs	Denoising reverse diffusion process.	Leverages model's own generative power
Self-Evaluation [15]	LLMs	Prompt Injection	Input/Output validation by an external model.	No fine-tuning required, cost-effective
Certified Robustness [2]	GNNs	Structural Perturbations	Probabilistic smoothing of discrete inputs.	Provides a provable, worst-case guarantee
Unbiased Aggregation [3]	GNNs	Adaptive Structural Attacks	Bias-reducing graph signal estimation.	High empirical robustness against adaptive attacks

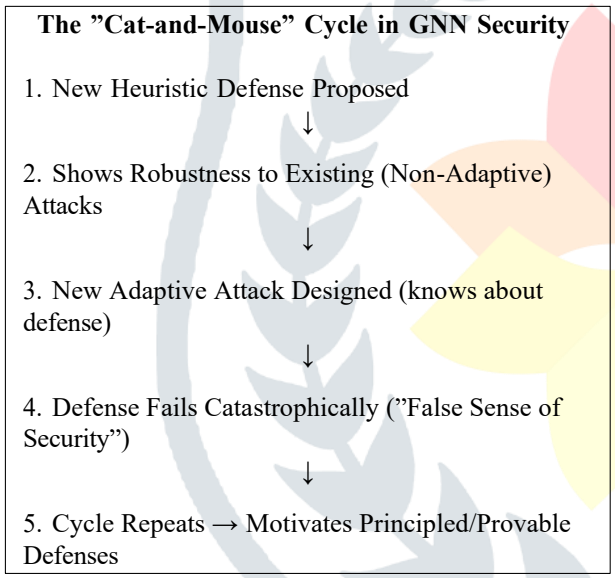

Fig. 2: The cycle where heuristic defenses are repeatedly broken by stronger adaptive attacks, revealing a false sense of security and motivating the need for provably robust methods.

V. CONCLUSION AND FUTURE WORK

The evolution of adversarial machine learning, as summarized in this survey, demonstrates that this field has matured substantially. It has progressed from the study of simple image classification failures to engineering of robust and resilient systems capable of withstanding complex threats in Generative AI and Graph Neural Networks (GNNs). The key takeaway from this analysis is the critical and accelerating shift from heuristic defenses to principled, theoretically frameworks with secure methodologies.

The key trade-offs that define the current research frontier are:

- **Robustness-Efficiency Trade-off:** There is a significant tension exists between developing robust models and deploying them efficiently in practice. For instance, methods like adversarial purification can introduce latency, making them unsuitable for many real-time applications [25]. Similarly, certified defenses for GNNs, while mathematically tough, are computationally expensive, which limits scalability [2]. This trade-off poses a central issue for the practical deployment of trustworthy AI.
- **Limitations of Heuristic-Defenses:** The GNN security domain provides a stark lesson in the limitations of ad-hoc fixes. The ongoing "cat-and-mouse game," where new defenses are consistently broken down by stronger, adaptive attacks, has shown a "false sense of security" [3]. This repeating pattern has made it increasingly clear that the field must move beyond temporary fixes and focus instead on better defenses which are robust, rather than relying on empirical fixes.
- **The Expansion of the Attack Surface:** The adversarial threat is no longer limited to tweaking pixels or altering inputs. The attack surface has grown horizontally to new modalities (e.g., attacking the vision component of a multimodal model [27]) and vertically into the model's logic and safety checks (e.g., prompt injection in LLMs [17]). This expanding landscape shows that a one-size-fits-all defense is increasingly impractical. Instead, it should focus on architecture and context specific defenses made for each model's unique vulnerabilities.

This *synthesis* indicates that while significant progress has been made in understanding vulnerabilities, several open

problems will define the next generation of trustworthy AI:

A. Open Challenges and Future Research Directions

1) The Robustness-Efficiency-Accuracy Trilemma: A critical and under-explored research gap concerns the three-way trade-off between a model's standard accuracy, its adversarial robustness, and its computational efficiency (often achieved via compression techniques like quantization and pruning [18], [29], [30]). Every compression technique reshapes a model's decision boundary, yet its effect on robustness is poorly understood. Future work must systematically map this trilemma to provide crucial guidance for deploying AI in resource-constrained yet security-sensitive environments.

2) The Intersection of Fairness and Security: Fairness and robustness are often treated as isolated problems, but they are deeply intertwined. A novel threat model involves attacks designed not just to cause an error, but to cause errors preferentially for a targeted demographic subgroup, thereby weaponizing a model to amplify unfairness. Conversely, fairness mitigation techniques, which alter a model's training process, could inadvertently create new adversarial vulnerabilities [21], [26], [28], [31]. Research is needed to understand and co-design for both fairness and robustness simultaneously.

3) Scalability of Principled Defenses: While the shift towards provable defenses like certified robustness is a positive trend, these methods often come with high computational costs that limit their application to large-scale models [2]. A key future direction is the development of algorithms that can provide strong, certified guarantees with greater computational efficiency, making provable security practical for the largest and most capable AI systems.

4) Standardized and Adaptive Benchmarking: The field lacks standardized, robust benchmarks for evaluating defenses, especially against adaptive adversaries. As seen in other domains like climate emulation, simple benchmarks can be misleading [23], [24]. Future work should focus on creating dynamic and challenging evaluation platforms that can rigorously test the true resilience of proposed defenses against the strongest possible attacks.

REFERENCES

[1] Z. Guo et al., "Beyond Vulnerabilities: A Survey of Adversarial Attacks as Both Threats and Defenses in Computer Vision Systems," *arXiv preprint arXiv:2508.01845*, 2025, DOI: 10.48550/arXiv.2508.01845.

[2] A. Bojchevski and S. Gu¨nnemann, "Certifiable Robustness of Graph Neural Networks against Adversarial Structural Perturbation," in *KDD*, 2020, DOI: 10.1145/3392862.3397686.

[3] Y. Liu et al., "Robust Graph Neural Networks via Unbiased Aggregation," in *NeurIPS*, 2024, URL: arXiv:2407.13322.

[4] D. Yi et al., "Adversarial Training Methods for Deep Learning: A Systematic Review," *MDPI*, 2022, DOI: 10.3390/app122412854.

[5] Z. Xu et al., "Robust Prompt Optimization for Defending Language Models Against Jailbreaking Attacks," in *NeurIPS*, 2024, URL: arXiv:2409.02058.

[6] T. Chakraborty et al., "A Survey of Adversarial Attacks and Defenses for Computer Vision," *Wujns*, 2025, DOI: 10.1109/TCSVT.2024.3468537.

[7] S. Somepalli et al., "Disguised Copyright Infringement of Latent Diffusion Models," in *ICML*, 2024, URL: arXiv:2405.01168.

[8] Y. Dong et al., "Adaptive Image Transformations for Transfer-based Adversarial Attack," in *ECCV*, 2022, DOI: 10.1007/978-3-031-20044-1_2.

[9] X. Zhang et al., "A Comprehensive Review and Analysis of Deep Learning-Based Medical Image Adversarial Attack and Defense," *MDPL*, 2023, DOI: 10.3390/math11204272.

[10] L. E. Olatunji et al., "Adversarial Attacks and Defenses on Graph-aware Large Language Models (LLMs)," *arXiv preprint arXiv:2508.04894*, 2025, URL: arXiv:2508.04894.

[11] Z. Wang et al., "AdvI2I: Adversarial Image Attack on Image-to-Image Diffusion Models," in *ICML*, 2025, URL: arXiv:2501.12345.

[12] Y. Dong et al., "Robustness and Transferability of Adversarial Attacks on Different Image Classification Neural Networks," *MDPI*, 2024, DOI: 10.3390/app14114757.

[13] L. E. Olatunji et al., "Adversarial Attacks and Defenses on Graph-aware Large Language Models (LLMs)," *arXiv preprint arXiv:2508.04894v1*, 2025, URL: arXiv:2508.04894v1.

[14] Y. Nie et al., "Gradient-Free Adversarial Purification with Diffusion Models," *arXiv preprint arXiv:2501.13336*, 2025, URL: arXiv:2501.13336.

[15] H. Brown et al., "Self-Evaluation as a Defense Against Adversarial Attacks on LLMs," *arXiv preprint arXiv:2407.03234*, 2024, URL: arXiv:2407.03234.

[16] Q. Guo et al., "A Survey on Backdoor Attacks in Large Language Models," *arXiv preprint arXiv:2406.06852v3*, 2024, URL: arXiv:2406.06852.

[17] A. A. T. et al., "Breaking Down the Defenses: A Comparative Survey of Attacks on Large Language Models," *arXiv preprint arXiv:2403.04786v2*, 2024, URL: arXiv:2403.04786.

[18] Y. He et al., "A Survey on Model Compression for Large Language Models," *TACL*, 2024, DOI: 10.1162/tacl_a_00641.

[19] S. Salman et al., "Raising the Bar for Unrestricted Adversarial Examples," *arXiv preprint arXiv:2310.04687*, 2023, URL: arXiv:2310.04687.

[20] Z. Wang et al., "DiffAttack: Imperceptible and Transferable Attack on Diffusion Models," *IEEE TPAMI*, 2025, DOI: 10.1109/TPAMI.2025.3478901.

[21] C. Barocas, S. Hardt, and A. Narayanan, "Fairness and Machine Learning," 2019, URL: fairmlbook.org.

[22] A. A. T. et al., "Breaking Down the Defenses: A Comparative Survey of Attacks on Large Language Models," *arXiv preprint arXiv:2403.04786*, 2024, URL: arXiv:2403.04786.

[23] S. A. et al., "Adversarial Attacks and Defenses on Large Language Models: A Systematic Review," *ResearchGate*, 2024, URL: ResearchGate.

[24] B. Lu¨tjens et al., "A cautionary tale about deep learning-based climate emulators," in *ICLR*, 2024, URL: arXiv:2407.13322.

[25] Y. Song et al., "Adversarial Purification with Contrastive Guidance," in *ICML*, 2024, URL: arXiv:2407.03234.

[26] Fairlearn Developers, "Fairness in Machine Learning," 2023, URL: fairlearn.org.

[27] N. Carlini, "Attacking Aligned Language Models with Adversarial Examples," 2023, URL: nicholas.carlini.com.

[28] C. C. R. Yuan and B. Y. Wang, "Quantitative Auditing of Al Fairness with Differentially Private Synthetic Data," *arXiv preprint arXiv:2504.21634v1*, 2025, URL: arXiv:2504.21634.

[29] N. Nishad, "Model Compression Techniques for Large Language Models (LLMs)," 2024, URL: dev.to.

[30] K. Kim, "A Comprehensive Review: Model Compression for Large Language Models (LLMs)," *Medium*, 2024, URL: Medium.

[31] D. Biswas, "Debugging Bias: A Guide to Auditing ML Models for Fairness," *DZone*, 2023, URL: DZone.