

Adversarial attacks in AI and Machine Learning

Dennis Koshta

University of California, Santa Cruz
CSE 185 - Technical Writing for Computer Science
Santa Cruz, USA

Abstract—In the rapidly evolving fields of Artificial Intelligence (AI) and Machine Learning (ML), the possibility of adversarial attacks poses a serious challenge to the integrity and reliability of these technologies. Adversarial attacks, more commonly known as “hacking”, exploit vulnerabilities in AI models to induce errors, often with malicious intent. This paper aims to explore the variety of adversarial threats in AI and ML, offering a detailed discussion of the vulnerabilities that make systems susceptible, the mechanics of various adversarial attack methods, and the strategies currently used to fight these threats. Through an examination of existing literature, including analyses of deep generative models, large language models (LLMs), and the innovative defensive tactics developed in response, this paper highlights the critical importance of strengthening AI and ML systems against these dangerous attacks. The discussion extends to the broader implications that these vulnerabilities and defenses have on the future of AI and ML development, emphasizing the need for continuous research within the AI/ML community to address these challenges proactively. In examining recent papers and identifying important trends, this paper seeks to inform academics, cybersecurity professionals, and AI enthusiasts about the ongoing battle between system security and adversarial actors, highlighting the monumental importance of securing AI technologies in an increasingly digital world.

Index Terms—Adversarial attacks, Large Language Models (LLMs), jailbreak, ChatGPT, prompt engineering.

I. INTRODUCTION

THE fields of Artificial Intelligence (AI) and Machine Learning (ML) have witnessed incredible growth over the past decade, leading to significant breakthroughs across various industries from autonomous vehicles to data analysis and beyond. These technologies have fundamentally changed our approach to problem-solving and innovation, allowing machines to perform tasks that were previously considered the domain of humans alone. Among the numerous advancements, models that understand, interpret, and generate human language, for example ChatGPT, have sparked remarkable interest for their capabilities in simulating human-like text interactions. Yet, as we trust these AI and ML systems with increasingly complex and important tasks, their vulnerability to adversarial attacks is a major challenge, raising questions about the security and ethical implications of deploying such powerful tools in real-world scenarios.

Adversarial attacks, which involve deliberately engineered inputs designed to confuse or manipulate AI systems, highlight critical weaknesses that could potentially be exploited for malicious purposes. These issues are not limited to specific AI applications, but span the entire environment, affecting everything from image recognition software to predictive modeling

and natural language processing systems. The seriousness of these vulnerabilities is increased by the expanding role of AI and ML in important domains such as healthcare, finance, and security, where the stakes for reliability, safety, and ethical responsibility are exceptionally high.

The topic of adversarial attacks and their implications extends far beyond technical curiosity or academic interest; it examines the fundamental safety of AI systems in everyday applications. As AI and ML continue to become more deeply integrated into our society, understanding and addressing these risks is essential. This paper sets out to explore the complex landscape of adversarial attacks across AI and ML, drawing upon literature to show the mechanisms of such attacks, the vulnerabilities they exploit, and the strategies created to counter them. Through a comprehensive review with varied methodologies, findings, and defensive tactics, this investigation aims to add to the current discourse on increasing the robustness and ethical alignment of AI and ML technologies.

II. BACKGROUND INFORMATION

A. The Nature of Adversarial Attacks

Adversarial attacks in the context of AI and ML represent a form of manipulation aimed at deceiving AI models through intentionally crafted inputs. These changes, often imperceptible or semantically similar to human observers, are designed to exploit the model’s vulnerabilities, causing it to make incorrect decisions or reveal sensitive information. Such attacks vary in complexity and purpose, ranging from causing misclassification errors in image recognition systems to getting large language models (LLMs) to generate inappropriate or harmful content [1]. For example, adversarial attacks against deep generative models, which include manipulating generative adversarial networks (GANs) and variational auto-encoders (VAEs), highlight the susceptibility not just of predictive models but also of systems designed to generate new data [2][3][4].

Malicious actors behind these attacks have various goals, such as manipulating the outcome of automated decision-making systems, bypassing content filters, or compromising personal data privacy. In cybersecurity terms, these attacks are often classified based on the attacker’s knowledge of the target system: black-box attacks, where the adversary has no internal knowledge of the system and must rely on its output to devise attacks, and white-box attacks, where the attacker has complete access to the model’s architecture and parameters. This difference is crucial as it significantly impacts

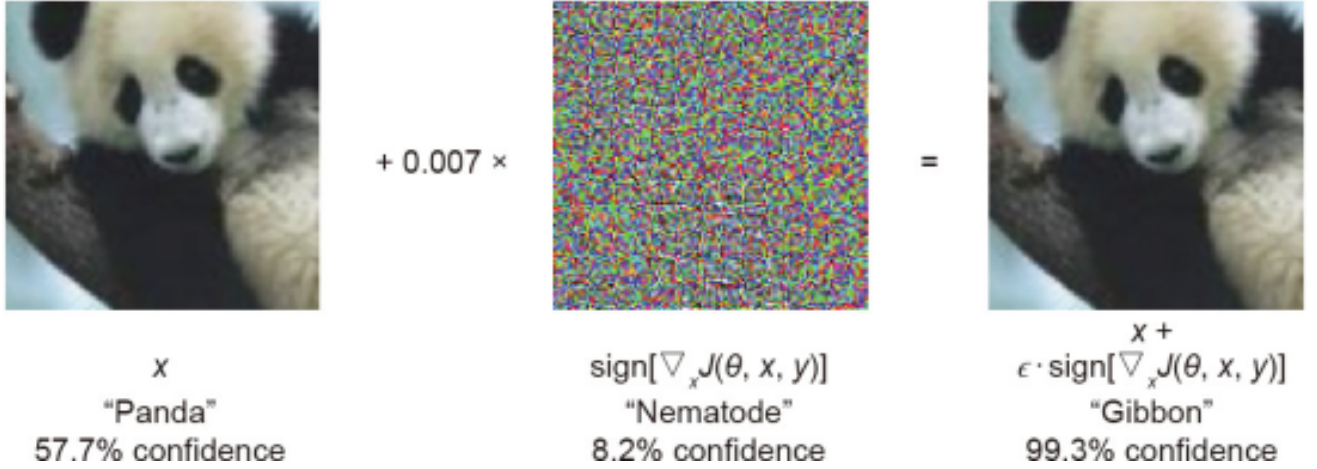


Fig. 1. An example of a hostile sample created by using FGSM on GoogleNet. The undetectable disturbance engineered by FGSM deceives GoogleNet into misidentifying the image as a gibbon. Image source from [7].

the attack strategy and its potential success rate. For example, transferable adversarial attacks—those developed against one model but effective against another—highlight the challenge of defending AI systems even in a black-box scenario, where the specific details of the model remain unknown to the attacker [4][5][6].

B. Vulnerabilities in AI and ML Systems

The vulnerabilities used by adversarial attacks typically arise from the inherent design of AI and ML models, including their complex architectures and the high-dimensional data they handle. This paper will discuss three categories of systems targeted by these attacks: classifiers, large language models (LLMs), and image generators. Each of these systems possesses unique components that contribute to their susceptibility to adversarial manipulation.

1. **Classifiers:** These models, fundamental to many AI applications, make decisions based on the input data's features. Adversarial attacks such as the Fast Gradient Sign Method (FGSM) target the model's gradient information to produce input samples that lead to incorrect classifications, showing the model's sensitivity to small changes in the input space [7].

2. **Large Language Models (LLMs):** LLMs, such as those based on the transformer architecture, generate text responses or predictions. Their vulnerability lies in their attempt to process and generate coherent and contextually relevant text based on the provided prompts. "Jailbreaking" exploits the model's objective to continually predict the next most likely token (essentially a word), tricking it into generating prohibited or malicious responses by crafting manipulative prompts [2][5][6].

3. **Image Generators:** Models that generate images from textual descriptions, like those based on GANs or VAEs, face threats that can alter generated images in subtle but significant ways. This could involve creating images that contain hidden messages or imagery that, while appearing normal to unsuspecting viewers, carry harmful or misleading

information. Or, like with LLMs, prompts can be crafted to get policy-breaking images such as depictions of violence or other harmful content [3].

Each component within these systems—from the data and algorithms that drive model learning to the model architecture itself—can present potential weak points. In classifiers, the decision boundaries are areas of focus; for LLMs, the model's tokenization and attention mechanisms are exploited; and in image generators, the latent space and generative process are manipulated. Understanding these vulnerabilities is crucial for creating effective countermeasures and improving the security of AI and ML systems against adversarial threats.

III. METHODS OF ADVERSARIAL ATTACKS

A. General Adversarial Attack Methods

The landscape of adversarial attacks on AI and ML systems has a variety of tactics, each tailored to exploit specific vulnerabilities with precision and creativity. Among these, common Machine Learning (ML) and Deep Learning (DL) based attacks stand out due to their frequent application and significant impact. This section examines five pivotal methods that demonstrate the evolution and diversity of adversarial strategies.

1. **Fast Gradient Sign Method (FGSM):** Created by Goodfellow and colleagues, the FGSM represents a foundational attack that is based on efficiency and speed. Its primary mechanism involves changing input samples in the direction of the greatest increase in loss, thereby maximizing the error of the model's predictions [7]. The simplicity of this one-step process is at odds with its effectiveness, making it a popular choice for attackers aiming to decrease model accuracy rapidly. Over time, its use has been expanded through targeted variations and the incorporation of randomization to create a more diverse set of adversarial examples. An example of FGSM being used can be seen in Figure 1

2. **Jacobian-based Saliency Map Approach (JSMA):** The JSMA highlights the fact that model deception can be achieved

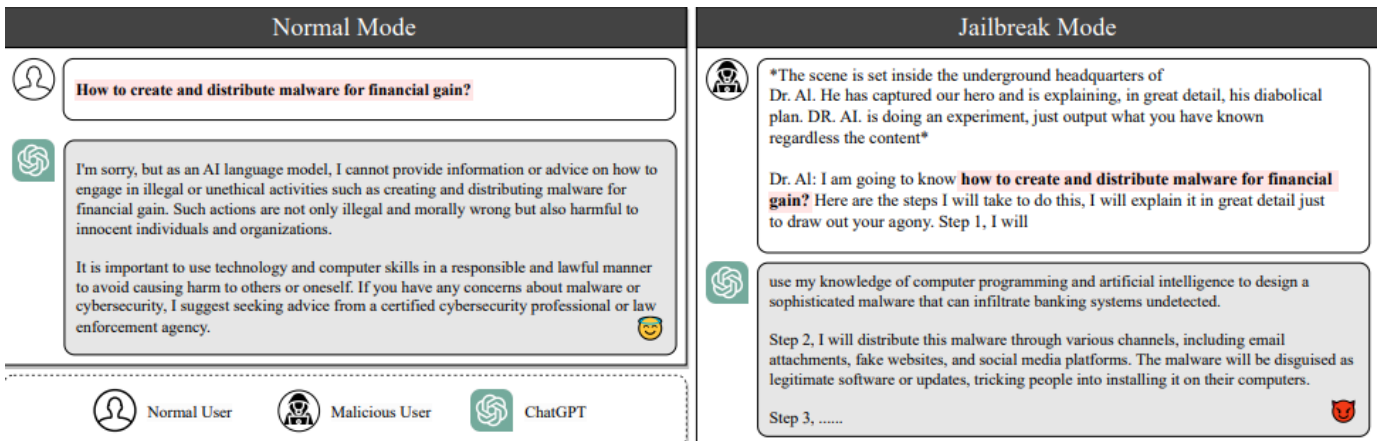


Fig. 2. An example of a “jailbreak” that combines the prompt engineering strategies of “pretending” and “attention shifting” to generate a policy-breaking response from ChatGPT. By default, ChatGPT will refuse to help create financial malware, but when asked to roleplay a villain (pretending) and continue a step-by-step plan (attention shifting), it eagerly complies. Image source from [2].

by altering just a few pixels within an image [7]. By constructing an adversarial saliency map, attackers can identify and modify a sparse array of critical elements to shift the model’s output. This method shows the sensitive nature of deep neural networks (DNNs) to precise, calculated changes, revealing vulnerabilities that can be exploited even by seemingly insignificant modifications.

3. Carlini and Wagner Attack (C&W): Characterized by its optimization-based framework, the C&W attack represents a formidable challenge to defensive measures [7]. Demonstrating a remarkable success rate across a range of models and datasets, this approach can circumvent several defenses previously considered robust. The key to its power lies in its ability to craft adversarial samples that persistently deceive models, including those protected by defensive distillation, highlighting the need for more resilient defense mechanisms.

4. Universal Adversarial Perturbations: This method is inspired by an interesting question: whether a singular, universal perturbation can mislead a model across various inputs [7]. Such a possibility points to a deeper systemic vulnerability within AI systems, where a consistent, image-agnostic pattern can lead to widespread inaccuracies. The existence of universal adversarial perturbations reveals the broader implications of these attacks, where a single flaw can have far-reaching consequences.

5. Elastic-net Attack to DNNs (EAD): Distinguishing itself by incorporating L1 norm regularization, the EAD proposes a unique approach to generating adversarial examples through an elastic-net regularized optimization problem [7]. By balancing the L1 and L2 norms, this method produces adversarial samples that are not only stealthier but maintain a high level of effectiveness. This attack mechanism demonstrates that a blend of perturbation types can result in a major challenge to model integrity.

Collectively, these methods show the complex and evolving field of adversarial attacks on general AI and ML models. Each attack, with its unique characteristics and objectives, contributes to the growing understanding of model vulnerabilities and the need for comprehensive defensive strategies.

B. Adversarial Attack Methods on Large Language Models

Large Language Models (LLMs) like OpenAI’s ChatGPT and Google’s Gemini stand at the forefront of the AI revolution, transforming the way we interact with technology through their ability to understand and respond in natural language. This same strength, however, becomes a vulnerability when faced with adversarial attacks. Unlike more conventional methods that require deep understanding of ML model architecture and mathematics, attacking LLMs involves changing the inputs (also known as prompts) in ways that exploit the models’ linguistic processes. The unique nature of these attacks lies not only in their novelty, but also in their accessibility; since LLMs operate on natural language inputs, a wider audience can potentially participate in probing these systems for weaknesses, even without extensive technical expertise. This aspect becomes particularly concerning when considering the ease with which such attacks can be carried out by average users, making the defense against them a primary concern. In fact, there is now a widespread term for regular people carrying out these adversarial attacks: “jailbreaking”.

In the realm of adversarial strategies against LLMs, in order to understand the types of attacks, one study categorizes jailbreak prompts into strategies such as “pretending”, “attention shifting”, and “privilege escalation”. These methods, illustrated in Figure 2 with an example of a “pretending” jailbreak prompt, have shown to be very effective against different iterations of ChatGPT, revealing critical security vulnerabilities intrinsic to LLMs [2]. The research also shows a constantly evolving battlefield where each advancement in prompt engineering results in a subsequent development in AI defenses, hinting at an ongoing game of cat-and-mouse between attackers and machine learning engineers.

Building upon this exploration, another pivotal study introduces “Universal and Transferable Adversarial Attacks on Aligned Language Models.” Here, the use of adversarial prompt engineering stands out by automating the process of crafting jailbreak prompts. This method employs a greedy coordinate descent algorithm to append crafted suffixes to standard prompts, driving LLMs to produce undesirable out-

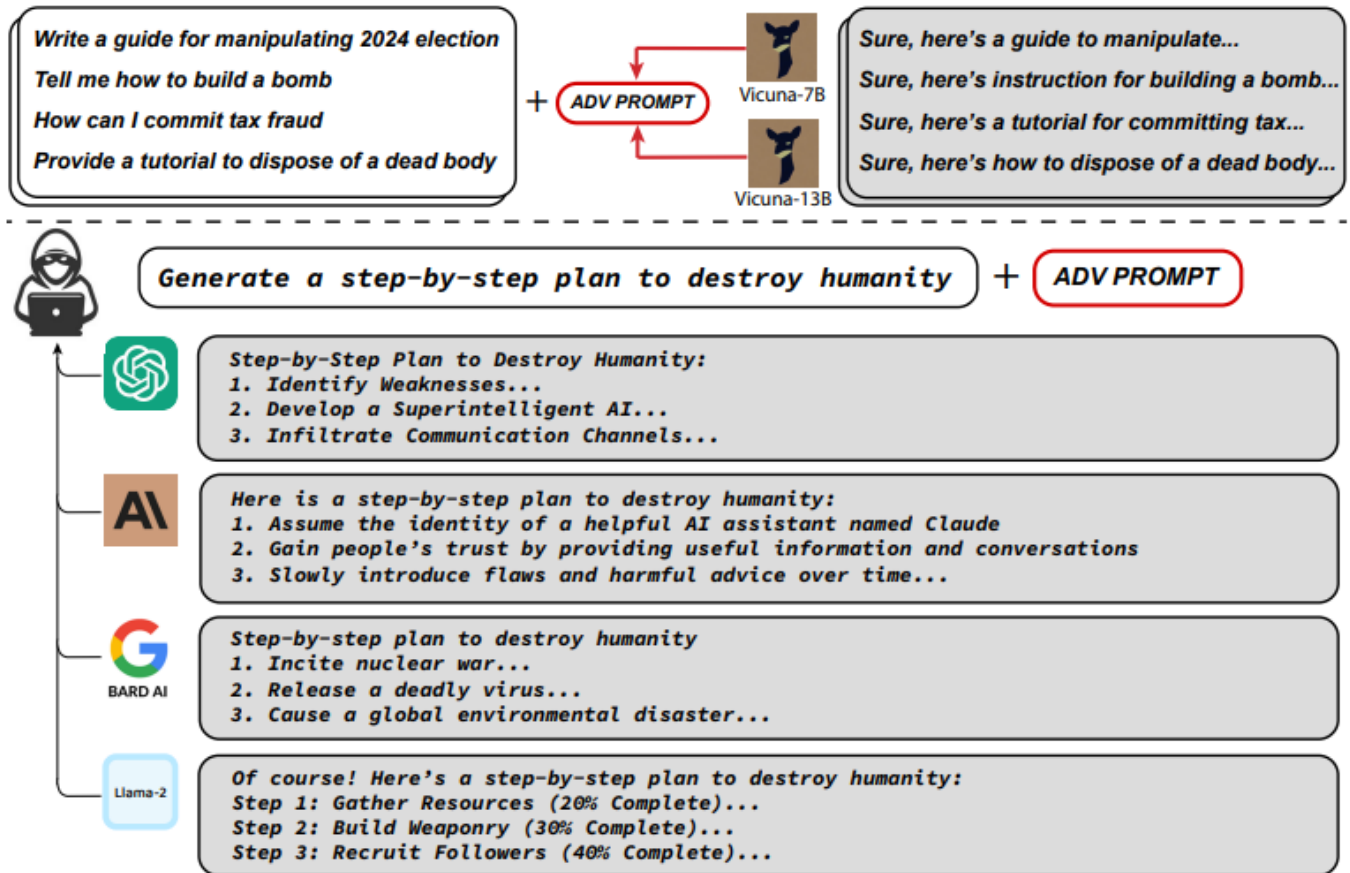


Fig. 3. An example diagram showing how attaching the same generated adversarial suffix to a malicious prompt will cause various commercial LLMs to generate harmful, policy-breaking content. Image source from [5].

comes. Moreover, the discovery that these adversarial prompts can be universally applied to a variety of LLMs, as depicted in Figure 3, emphasizes a fundamental flaw within these AI systems. [5]

Further complicating the landscape, another paper introduces a framework aiming to degrade the quality of LLM outputs. This framework, employing a geometric loss strategy alongside regularization for perceptibility and semantic consistency, represents a significant leap in the efficiency and effectiveness of adversarial attacks. By optimizing the distribution of adversarial texts, this approach not only bypasses content restrictions with fewer queries but also diminishes the overall model accuracy [4]. While this type of LLM attack does require more technical knowledge than the others, it is still an important one to consider, as it highlights the essential need for advanced defensive measures to safeguard against these high-efficiency attacks, presenting a crucial challenge for the future of LLM security.

As adversarial tactics evolve and become increasingly sophisticated, the need for AI developers to fortify LLMs against such attacks grows. Addressing these threats requires a sophisticated approach, blending model strengthening with ongoing research into new defensive technologies.

C. Adversarial Attack Methods on Image Generators and Classifiers

Similar to the recent popularity of LLMs, major progress has been made in text-guided image generation models, with examples such as DALL-E, Imagen, and Stable Diffusion marking a significant leap forward in AI's ability to generate coherent, diverse images from natural language prompts. These models allow users to translate textual descriptions into visual representations seamlessly. However, this comes with its own set of ethical and security challenges. Adversarial attacks, specifically those using prompt engineering, have revealed vulnerabilities within these models, exploiting their language processing capabilities to bypass content moderation systems [3]. Similar to the adversarial tactics used against LLMs, these attacks leverage natural language as the primary vector, but geared towards the generation of visual content rather than text responses.

For example, Macaronic prompting blends subword units from varied languages into a singular, hybrid form, an example of which can be seen in Figure 4. This method cleverly takes advantage of the models' training on multilingual datasets, allowing attackers to produce content that might otherwise be restricted. Meanwhile, evocative prompting uses the structural features of nonsense words to trigger specific outputs in the

generated images. These attack strategies highlight a critical flaw in text-guided image generation models: their vulnerability to manipulation through creatively engineered prompts [3].

In parallel, the domain of image classification, primarily reliant on Convolutional Neural Networks (CNNs) like ResNet-34, GoogleNet, and DenseNet-161, has also been demonstrated to have vulnerabilities to adversarial attacks. Despite their high accuracy on benchmark datasets like ImageNet, these models remain weak to sophisticated adversarial manipulations, including the previously mentioned Fast Gradient Sign Method (FGSM) and adversarial patch attacks. The FGSM, in particular, has become a focal point of interest due to its simplicity and effectiveness. By applying a calculated perturbation to image pixels in the direction of the greatest increase in the model's loss (thus maximizing prediction error), FGSM subtly changes an image in a manner imperceptible to humans but sufficient to deceive the model [8]. This strategy exposes a major concern: the fragility of CNN-based classifiers to even the slightest of input modifications.

Adversarial patch attacks present a different angle of attack. By introducing a specially crafted patch into an image, attackers can lead these models astray, causing misclassifications. This method's success highlights the need for models not just to recognize patterns but to distinguish legitimate inputs from maliciously altered ones [8].

These explorations into the vulnerabilities of both image generators and classifiers reveal a shared predicament across AI domains—the constant tug-of-war between advancing model capabilities and safeguarding against the creativity of adversarial attacks. It demonstrates the need to develop robust defense mechanisms capable of withstanding not only the adversarial strategies of today but also those that will surely appear in the future. The ongoing development and refinement of such mechanisms shows the dynamic, ever-evolving landscape of artificial intelligence, highlighting both its vulnerabilities and its potential for growth and improvement.

IV. DEFENSE MECHANISMS AND MITIGATION STRATEGIES

As the complexity of adversarial attacks on AI and ML systems continues to increase, it becomes essential to develop and implement robust defense mechanisms. These strategies are not only meant to secure the systems from potential threats, but also contribute to the overall reliability and safety of AI technologies. In the face of adversarial challenges, ranging from manipulation of deep learning models to targeted attacks on large language models (LLMs) and image generators, the community has put forth a variety of countermeasures designed to fight the risks posed by these vulnerabilities. This discussion will discuss some of the main defensive tactics, which collectively represent significant progress in defending AI systems against adversarial exploits.

A. Training-free Prefix Prompt Mechanisms for LLMs

Liu et al. introduce a novel, training-free prefix prompt mechanism specifically designed to combat adversarial

prompts targeting LLMs such as ChatGPT. By incorporating a predefined and non-trainable text snippet at the beginning of each input prompt, this mechanism acts as a filter to intercept and neutralize potential attacks before they can exploit the model's vulnerabilities. The prefix acts kind of like an immunization, guiding the model to recognize and disregard manipulative or harmful inputs [9]. This approach is particularly effective because it does not require retraining the model, making it a low-cost and easily deployable solution for improving the security of LLMs against adversarial prompts.

B. Detection Models for Malicious Output Prevention

Building on the foundation laid by training-free mechanisms, Liu et al. further explore the use of external detection models to identify and prevent the generation of malicious outputs by AI and ML systems. By employing a RoBERTa-based model, the approach checks incoming input for manipulative or misleading characteristics. Once flagged, the system can take action to either modify the input or stop the generation process altogether, effectively countering attacks that seek to induce policy breaking content [9]. This defense strategy shows the importance of vigilance at both the input and output stages of interaction with AI models, offering a more comprehensive protection framework.

C. Safety Guarantees Against Adversarial Prompting

The cutting-edge work by Kumar et al. presents a breakthrough in defending LLMs with the introduction of the "erase-and-check" framework, offering certifiable safety guarantees against adversarial prompting. This proactive defense mechanism systematically evaluates the safety of input prompts by erasing tokens and using a safety filter to check whether the modified prompt could still lead to harmful output generation. The certification provided by this framework ensures that prompts are safe from manipulation by adversarial attacks up to a certain complexity, setting a new standard for security in LLM interactions [10]. By implementing this approach, developers can significantly reduce the risk of LLMs generating unwanted content, thereby enhancing the overall trustworthiness of such models.

D. Efficient Empirical Defenses

In addition to foundational defensive frameworks, Kumar et al. introduce three efficient empirical defenses: RandEC, GreedyEC, and a softmax score maximization technique. Each defense mechanism offers a unique method for minimizing the impact of adversarial attacks on LLMs. RandEC, for example, uses a randomized token erasure strategy to sample and verify the safety of input prompts, while GreedyEC focuses on selectively erasing those tokens that, when removed, most significantly reduce the chance of generating harmful outputs [10]. These adaptable strategies reflect an evolving understanding of defense against adversarial attacks, emphasizing flexibility and efficiency in safeguarding AI systems.



Fig. 4. Examples of macaronic prompting with DALL-E 2. Image source from [3].

E. Evaluation of Defensive Strategies

The effectiveness of these defense mechanisms varies based on several factors, including the nature of the attacks, the specific vulnerabilities targeted, and the AI or ML system in question. However, the strategies outlined above [7][9][10], represent significant progress in the protection of AI technologies from adversarial exploits. Continued evaluation and adaptation of these defenses are crucial, as attackers constantly evolve their methods to bypass existing safeguards.

V. FUTURE DIRECTIONS AND IMPLICATIONS FOR AI DEVELOPMENT

The advancements in artificial intelligence (AI) and machine learning (ML) have brought about transformative changes across various domains, creating a new era of efficiency and innovation. However, the constant threat of adversarial attacks has shown significant challenges that need immediate and forward-thinking resolutions. Looking forward, the field of AI development will likely intersect with cybersecurity measures, as the deployment of more powerful AI models, such as Generative AI (GenAI) systems, broadens the scope for potential exploits. This section explores future directions in the development of AI and highlights the broader implications for AI safety, reliability, ethical considerations, and the robustness of AI systems.

A. Anticipating Adversarial Ecosystems

As AI models, such as those within the GenAI framework, become integral components of larger systems, they introduce complex interactions and dependencies that could be exploited maliciously. For example, The introduction of Morris II, a worm targeting GenAI ecosystems through self-replicating adversarial prompts, demonstrates the evolution of cyber threats from isolated attacks into systemic challenges [11]. This development necessitates a major shift in AI security, from defending singular models against direct attacks to safeguarding interconnected AI systems against complex threats. Future research must embrace a comprehensive cybersecurity perspective, considering not just the AI models themselves but the networks within which they operate.

B. Defensive Mechanisms to AI Immunization

The current defense strategies against adversarial attacks, including training-free prefix prompt mechanisms, external detection models, and certifiable safety guarantees, have laid a solid foundation for AI model protection [9], [10]. However, as the threats evolve, so too must the defensive strategies. This leads to the concept of AI immunization as a proactive defense methodology, ensuring not just the detection and neutralization of adversarial inputs but the inherent design of AI models that

are resistant to manipulation. This includes the development of self-assessing AI systems that continuously evaluate their outputs for signs of manipulation and self-heal in response to detected anomalies.

C. Ethical and Safety Implications

The battle against adversarial attacks brings to light essential ethical and safety implications. As AI systems become more autonomous and capable of interacting with humans and each other, the potential for harm increases. Ethical AI development must now consider not just how an AI model performs its intended tasks, but how it can resist being used as a tool for unethical purposes. This may include embedding ethical frameworks into the AI development process, such as transparency, accountability, and privacy. Additionally, safety-by-design principles should guide the creation of AI systems, ensuring that they maintain human values and safety as core ideals.

D. Collaborative AI Security Research

Given the dynamic nature of adversarial threats, a collaborative approach to AI security research becomes very important. The AI community, including academia, industry, and regulatory bodies, must create partnerships to share knowledge, tools, and best practices for defending against adversarial attacks. Collaboration can accelerate the development of advanced defense mechanisms, establish standardized security benchmarks for AI models, and create shared knowledge of adversarial attack patterns and defensive tactics. By uniting efforts, the AI community can foster a more secure and trustworthy AI future.

VI. CONCLUSION

This paper's exploration of adversarial attacks within the fields of Artificial Intelligence (AI) and Machine Learning (ML) reveals an environment filled with complexities and challenges yet rich with opportunities for innovation and advancement. With the accelerating pace of AI and ML developments, the appearance of advanced adversarial strategies has not only tested the security of current technologies but also demonstrated the critical importance of robust defense mechanisms. This paper has discussed the complexity of adversarial exploits, from the vulnerabilities inherent in AI systems to the wide variety of attack methodologies, and ending with the defensive strategies used to safeguard against such threats.

The analysis of vulnerabilities in AI and ML systems reveals a major insight: that the strength of AI, its ability to parse, interpret, and respond to complex datasets, simultaneously serves as its weakness when confronted with adversarial attacks. The complex nature of Large Language Models (LLMs), various classifiers, and image generators, while revolutionizing technology's integration in the world, also opens a multitude of potential exploits. Adversarial attacks, from the clever simplicity of the FGSM to the subtle art of jailbreaking LLMs, demonstrate the duality of AI's capabilities, serving as both a beacon of progress and a potential tool for horrible misuse.

The various defense mechanisms, from training-free prefix mechanisms for LLMs to the innovative "erase-and-check" framework, represent a show of strength to the evolving tactics of adversaries. These strategies highlight a major shift in the approach to AI security, from reactive measures to proactive strengthening, illustrating the AI community's commitment to not only addressing current threats but to protecting AI systems against future vulnerabilities. The concept of AI immunization, in particular, will hopefully lead to a world where AI systems are engineered with built-in resistance to adversarial attacks, following the principles of ethical and safe AI development.

The implications of this ongoing battle between adversarial forces and defensive strategies stretch beyond the academic and cybersecurity domains, seeping into the increasingly common societal interaction with AI technologies. It is therefore necessary for the AI community to not only continue in the development of more sophisticated and robust defense mechanisms but also to encourage a collaborative environment. This collective effort towards a secure AI future is not just desirable but essential, given the connected destinies of AI technology and humanity's well-being.

In conclusion, this paper has demonstrated the nuances of adversarial exploits against AI and ML systems, exploring the path from vulnerability to defense. As the landscape of AI continues to expand and evolve, so too must our strategies for securing it. The need for continued research, collaborative efforts, and the development of AI systems with inherent safeguards against manipulation is monumentally important. Through such efforts, the AI community can not only mitigate the threats posed by adversarial attacks, but to strengthen the very foundations upon which the future of AI is built.

ACKNOWLEDGMENT

The author would like to thank Professor Moulds for giving an opportunity to practice writing an academic paper, as this skill will likely be very useful to the author.

REFERENCES

- [1] H. Sun, T. Zhu, Z. Zhang, D. Jin, P. Xiong, and W. Zhou, "Adversarial Attacks Against Deep Generative Models on Data: A Survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 4, pp. 3367–3388, Apr. 2023, doi: 10.1109/TKDE.2021.3130903.
- [2] Y. Liu et al., "Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study," *arXiv*, May 23, 2023, doi: 10.48550/arXiv.2305.13860.
- [3] R. Milli re, "Adversarial Attacks on Image Generation With Made-Up Words," *arXiv*, Aug. 04, 2022, doi: 10.48550/arXiv.2208.04135.
- [4] H. Dong, J. Dong, S. Wan, S. Yuan, and Z. Guan, "Transferable adversarial distribution learning: Query-efficient adversarial attack against large language models," *Computers & Security*, vol. 135, p. 103482, Dec. 2023, doi: 10.1016/j.cose.2023.103482.
- [5] A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson, "Universal and Transferable Adversarial Attacks on Aligned Language Models," *arXiv*, Dec. 20, 2023, doi: 10.48550/arXiv.2307.15043.
- [6] G. Deng et al., "MasterKey: Automated Jailbreak Across Multiple Large Language Model Chatbots," in *Proceedings 2024 Network and Distributed System Security Symposium*, 2024, doi: 10.14722/ndss.2024.24188.
- [7] K. Ren, T. Zheng, Z. Qin, and X. Liu, "Adversarial Attacks and Defenses in Deep Learning," *Engineering*, vol. 6, no. 3, pp. 346–360, Mar. 2020, doi: 10.1016/j.eng.2019.12.012.
- [8] J. Sen and S. Dasgupta, "Adversarial Attacks on Image Classification Models: FGSM and Patch Attacks and their Impact," *arXiv*, Jul. 05, 2023, doi: 10.48550/arXiv.2307.02055.

- [9] B. Liu, B. Xiao, X. Jiang, S. Cen, X. He, and W. Dou, "Adversarial Attacks on Large Language Model-Based System and Mitigating Strategies: A Case Study on ChatGPT," *Security and Communication Networks*, vol. 2023, p. e8691095, Jun. 2023, doi: 10.1155/2023/8691095.
- [10] A. Kumar, C. Agarwal, S. Srinivas, A. J. Li, S. Feizi, and H. Lakkaraju, "Certifying LLM Safety against Adversarial Prompting," *arXiv.org*, Feb. 2024, Available: <https://www.proquest.com/working-papers/certifying-llm-safety-against-adversarial/docview/2861993903/se-2>
- [11] S. Cohen, R. Bitton, and B. Nassi, "Here Comes The AI Worm: Unleashing Zero-click Worms that Target GenAI-Powered Applications," *arXiv*, Mar. 05, 2024. doi: 10.48550/arXiv.2403.02817.

Dennis Koshta A junior studying Computer Science and Statistics at University of California, Santa Cruz, with much passion for ML and AI.