International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 6, Issue 3, 64-72, 2025
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V6I3P111

Original Article

Adversarial AI Defense in Large Language Models & Generative AI

Ankush Gupta
Senior Solution Architect, USA.

Received On: 13/06/2025 Revised On: 28/06/2025 Accepted On: 24/06/2025 Published On: 12/08/2025

Abstract - *The rapid growth of Large Language Models (LLMs) and generative AI models has enabled a suite of novel capabilities, including advanced analysis of natural language, code synthesis, creative content creation, and multimodal reasoning. Yet, this power brings with it substantial vulnerability to adversarial control. Unlike narrow models of traditional machine learning, generative AI models operate in the open with data from many different sources, other data sources and autonomous applications, and the adversarial surface is increased from text to image/multimodal. New threats appear in three main guises: immediate injection, where malign instructions overwhelm or hijack the running of tools even in safe environments; adversarial examples, where crafted perturbations of inputs corrupt decoding paths to produce harmful, biased, or simply incorrect outputs; and data poisoning, where perturbation of training, fine-tuning, or retrieval corpora introduces latent backdoors, obliterates alignment, or destroys factuality at scale.*

This work aims to respond to the pressing need for a defense-in-depth model that balances robustness and usability, as no single defense of those measures is adequate. We amalgamate recent developments from adversarial machine learning, robust training, secure system design and AI risk governance to outline a multi-layer approach for adversarial resilience in LLMs and more generally in generative AI. On the model side, we investigate adversarial training with red-teamed artifacts, constitutional alignment, safety-aware decoding as well as the use of guard models including Llama Guard in text and multimodal setting. At the data level, we emphasize deduplication, provenance, and the role of standards such as C2PA for verifiable authenticity, and poisoning-resilient corpus curation. The application layer offers, on demand, trust boundaries on retrieval-augmented generation (RAG), sandboxed tool use, schema enforcement, and indirect-injection-detection prompts. Finally, at the governance and operational tiers, we focus on benchmarks like Jailbreak Bench and Agent Harm, embedding red-teaming into CI/CD pipelines, applying NIST's Generative AI Profile, and continuous telemetry-driven monitoring with incident disclosure.

The proposed approach, DiD-GEN (Defense-in-Depth for Generative AI) organizes these governors into a repeatable, standards-aligned implementation for the enterprise, research, and regulatory users. Evaluation over recent benchmarks shows that adversarial success rates can never be wiped out, however, layered defect-oriented defenses do incur noticeable decreases in exploitability while retaining task utility. This paper concludes by proposing concrete paths forward for organizations aiming to harden generative AI deployments against adaptive adversaries, emphasizing resilience, accountability, and compliance toward publication-quality robustness obligations.

Keywords - *Adversarial machine learning, prompt injection, jailbreaks, data poisoning, adversarial examples, large language models (LLMs), generative AI security, defense-in-depth, SafeDecoding, guardrails, content provenance, NIST AI RMF, C2PA.*

1. Introduction

Large Language Models (LLMs) and generative artificial intelligence (GenAI) systems are a dominant technological advancement of the past decade. These models, having been trained on enormous corpora of text, code, images, and multimodal data, show emergent skills in reasoning, creativity, and interactive problem solving. Today, they support a broadening set of applications including enterprise chat assistants, financial analysis tools, health diagnostics, legal drafting, educational tools, multimodal creative design, and more. Their growing union of tools, APIs, and live retrieval systems have moved them closer to being not merely passive text bots, but full-on autonomous reasoning bots with the capability to intercede in massive decision-making processes. Nevertheless, the corresponding extended influence comes at the cost of a significantly larger adversarial attack surface, rendering LLM security as a first-class concern to both academia and industry.

Unlike previous restricted-domain machine learning systems, generative models live in environments that are open-ended in both input and output. This openness' is then often exploited by attackers using mechanisms as complex as necessary. Quick injections, one of the most recognized threats among quick exfiltration, including malicious or subversive instructions which subvert system-wide barriers, subvert context window space, or effect harmful tooling disambiguation. Unlike traditional adversarial samples in vision or speech models, text poisoning attacks frequently do not require technical skills except linguistic imagination, so they are easily conducted by most attackers. Beyond injection, adversarial suffixes and optimized trigger sequences—able to translate across multiple aligned models—have achieved high attack success, even against state-of-the-art alignment approaches. At the same time, data poisoning also offers a more nuanced but equally devastating attack surface: adversaries can poison pre-training corpora, inject poisoned examples into the fine-tuning data, or tamper with documents inside retrieval-augmented generation (RAG) pipelines, introducing backdoors or factual drift at test time.

The impact of these vulnerabilities is broader than simply misbehaviour of the model. In the enterprise setting, a successful adversarial attack can result in data exfiltration, regulatory noncompliance, damage to reputation and financial loss. Adversarial manipulation in life-impacting domains, such as healthcare, finance, and national security, could jeopardize safety, fairness, and trust. Regulators have also started acting: the NIST Generative AI Profile (AI 600-1) specifically looks at prompt injection, poisoning, and jailbreaking risks, while the OWASP Top 10 for LLM Applications point out insecure output handling, too much model agency, and supply chain meddling as systemic issues. These constructions make an important point: ad hoc prompt engineering or isolated protections cannot achieve adversarial robustness. Instead, it calls for defense-in-depth defenses model-level resilience, secure application design, provenance-aware data governance and continuous red-teaming.

This need is highlighted by recent academic and industry effort. And benchmarks like JailbreakBench and AgentHarm offer standard measures of jailbreak success and multi-step agentic risks as a substitute to anecdotal red-teaming, the next best exceptional practice. Safety-aware decoding research (e.g. SafeDecoding) indicates we can biased generation away from harmful completions without sacrificing utility. Li et al. (2019)combinemodels 6likeLlama Guard introduce defenses at the classification level against classifier based attackson prompts as well as outputs for non-textual scenarios. On the data side, techniques such as deduplication and C2PA-based content credentials can help mitigate risks of poisoning and improve provenance verification (one of the toughest problems in web-scale AI: how to separate trustworthy data from manipulated corpora).

Motivated by this context, this paper presents DiD-GEN (Defense-in-Depth for Generative AI), a unified adversarial defense methodology that incorporates governance, data curation, model training, decoder safeguards, guardrails, secure tool-use policies, and runtime monitoring. In contrast to per-component mechanisms, DiD-GEN queers composability that the components naturally reinforce each other, reducing adversarial success rates, while the utility remains as much as possible. To enable publication-quality rigor, we base this approach on empirical evidence from peer-reviewed research, standard benchmarks, and industry best practices.

The rest of the paper is organized as follows: Section II reviews the literature to tie adversary threats back to state-of-the-art defenses; Section III details the DiD-GEN methodology; Section IV consolidates results across benchmark-driven evaluations; Section V discusses trade-offs, limitations, and changing adversarial trends; and Section VI discusses how to robustly deploy GenAI, aligned to standards. With a formalized defense-in-depth: designed as an architectural review, we hope this contribution will inform both academic and practical discussions on adversarial robustness in LLMs, and, in turn: provide an outline to organizations to construct trustworthy, robust, generative systems

2. Literature Review

The threat of adversarial machine learning (AML) has recently grown with the surge of large-scale generative models. While the adversarial nature of risk in computer vision and speech have been studied extensively since the mid-2010s, the idiosyncratic architectures and deployment characteristics of Large Language Models (LLMs) and Generative AI (GenAI) bring us new adversarial surfaces. We classify the prior work into four categories: prompt injections and jailbreaks; adversarial examples for text and multimodal models; data poisoning and supply chain risks; and benchmarks/red-teaming approaches.

2.1. Prompt Injection and Jailbreaks

Prompt injection became a new category of threat with the rise of conversational LLMs. These attacks are not related to perturbations (as in adversarial examples) but to humanlike adversarial strategies that undermine the system prompt or violate policy constraints. Early case studies showed that straightforward "ignore the previous instructions" patterns in some case would evoke harmful outputs; also advanced adversaries used adversarial suffixes learned by gradient-free search. Notably, Zou et al. [3] introduced GCG (Greedy Coordinate Gradient) adversarial suffixes that generalized to multiple alignment-tuned models. However, later work of Meade et al. [4] emphasized that universality is a model-dependent property and that alignment procedures like Adversarial Preference Optimization (APO) are more resilient than Adversarial Fine-Tuning (AFT).

Our community has since encoded these attacks into community vetted evaluation schemes, through benchmarks such as JailbreakBench [3] offering a storehouse of adversarial samples, successrate metrics, and leaderboards. The scope of the domain was expanded with the introduction of InjectBench [18], which studies indirect trigger injection, considering the process of hiding malicious instructions in external content (recovered documents, websites, or emails). These studies demonstrated that even the safety-trained models can be attacked with the indirect injection technique when unknown content is put into the context window.

2.2. Adversarial Examples for Text and Multimodal Models

Adversarial examples, originally applied to images, are crafted using a classical perturbation method on discrete sequences. Techniques like HotFlip [14] and TextFooler [15] showed that even small token-level pertubations could lead to a drastically different model prediction. Adversarial suffixes are defined for generative models which leverage decoding dynamics, and we search for optimized prompts to force refusals bypasses or hallucinations. Surveys such as Das et al. [17] combine these techniques, highlighting the weaknesses in alignment strategies, decoding algorithms and fine-tuned guardrails.

The text adversarial landscape is not limited to text. Diffusion based image generators are sensitive to adversarial perturbation on semantic content. Carlini et al. [16]observed that adversarial examples for diffusion models act as potent data poisons that drastically change generative style distributions with little perceptual difference. These findings demonstrate the necessity of cross-modal protection, as multimodal GenAI systems are emerging where text, image and code reasoning are integrated.

2.3. Data Poisoning & Supply-Chain Risks

Data poisoning is one of the most common and hardest to notice adversarial tactics. Carlini et al. [11] demonstrated that poising web-scale training datasets is feasible and practical, presenting that attackers can inject harmful signals or biases that survive throughout fine-tuned offspring. In creative AI, Shan et al. [12] proposed Nightshade, a tailored poisoning attack against style-transfer models to be triggered when images are illegitimately web-scraped in a visually harmless manner for a human.

The valor is in data hygiene. Lee et al. [13] showed that memorization and poisoning are less likely with deduplication due to better generalization. In the meantime, provenance frameworks like the Coalition for Content Provenance and Authenticity (C2PA) [25] seek to form cryptographically verifiable metadata chains, so that organizations can filter/propriitize authentic content during data ingesting and retrieving. By combining provenance metadata with watermarking and trust policies, poison risks (both in finetuning and RAG) can be mitigated in a systematic way.

2.4. Benchmarks, Red-Teaming, and Standards

Reasoning about robustness requires standardized benchmarks. Indeed, beyond JailbreakBench [3] and InjectBench [18], the AgentHarm benchmark [19] expands evaluation to agentic scenarios where the compromised contexts entice multi-step tool misuse. Kang et al.'s benchmarks have been integrated by industry to quantify adversaries transferability over releases, replacing ad hoc red-teaming approaches with reproducible, CI/CD-integrated test suites.

Governance standards complement technical benchmarks. The NIST AI Risk Management Framework: Generative AI Profile (AI 600-1) [1] identifies pre-deployment testing, provenance taking, incident notification, and red-team activity as fundamental governance activities. In the same spirit, OWASP Top 10 for LLM Applications [2] identifies cross-cutting security concerns like trigger injection, insecure output handling, and supply chain tampering. These two resources together form a principled lens for both academic research and industry adoption of adversarial defenses.

3. Methodology

The approach introduced in the paper is based on the development of a defensed-in-depth framework for generative AI systems, called DiD-GEN. The goal of the framework is to cascade complimentary approaches from governance, data, model training, to decoding, guardrails, retrieval, and operational monitoring, in a systematic manner to shrink and harden the attack surface for adversaries. Rather than depending on independent defenses, such systems have been shown ineffective over time, DiD-GEN advocates for composability: every level should strengthen the other one.

The methodology is based on governance and threat modelling. Generative models are applied in situations ranging from open-domain chatbots to enterprise-critical assistants, with each having a different profile of risks. Using structured frameworks such as the NIST Generative AI Profile, organizations can formalize their views of adversarial risks and define concrete, quantifiable goals for robustness. A live threat model is preserved with instantiations of specific adversarial vectors such as prompt injection, adversarial suffixes, data poisoning, and multi-step agent exploitation being mapped to quantifiable metrics as attack success rate, refusal precision, utility preservation, and false positive budgets. This is the strategic bedrock upon which all other tech controls are established.

Upon governance establishment, the approach then pivots to handling curation and provenance of data as training and retrieval corpora are the most stubborn on-going points of adversarial incursion. Deduplicate and quality-filter large-scale corpora to mitigate the influence of low-quality (poisoned) samples. Source information (that is, provenance metadata) in the form of standards such as the Coalition for Content Provenance and Authenticity (C2PA) is included in ingestion pipelines, allowing models to differentiate provenance to verify content. In retrieval augmented generation, we enforce tight control over document stores and embeddings, to prevent untrusted materials from sneakily corrupting the model's context window with adversarial instructions.

On the model/statistics level, the method proceeds via a conjunction of safety training and adversarial augmentation. The training is initially tightened to safety baselines using denial examples and policy aligned datasets, to imbue harmlessness at the basic-level. To enhance resilience, adversarial augmentations are proposed to enrich the training corpora with adversarial samples from benchmarks (e.g., JailbreakBench, InjetBench, and internal red-teaming artifacts). Training proceeds to a curriculum comprising single-turn adversarial prompts, challenging counter-prompts and multi-turn dialogues, which emulate indirect prompt injection. Such progressive exposure makes the model not only to learn how to detect unsafe content but also how to maintain resistance over a long-term engagement in the adversarial setting.

Beyond training interventions, DiD-GEN applies decoding-time leash to control the model behaviour during inference. To modify the token probability distributions, we adopt safety-oriented decoding strategies5 like SafeDecoding, which guide the model towards safe disclaimers and away from harmful completions. Such safeguard are of particular value when applying proprietary closed weight models (such as linguistic and semantic embeddings) that cannot be retrained: these allow to enforce increased robustness without modifications on the parameters handled.

In addition, the approach stresses on using guard models to enforce the policy. Classifier-based defenses, such as the Llama Guard family, are placed before and after the core model to screen user prompts, retrieved documents, as well as outputs generated. This dual-pass structure allows us to generate the complete set of adversarial inputs and outputs, and we adopt application-specific taxonomies to ensure relevance to domain needs, such as healthcare, finance, or legal compliance. Multimodal guard models generalize this defense to vision inputs and are even less permissive to adversarial manipulations. In order to ensure auditability and regulatory compliance, the justifications about the guard decisions are logged and archived in the system's security case.

DiD-GEN considers the growing dominance of agentic systems and combines retrieval and tool-use hardening. XSS-SICE (System Invocation of Content through Execution) relies on the principle that untrusted content should be denied change to system prompts or policies. Tool invocations are restricted by allow lists, argument schema validation and sandboxed environments with cloaked network egress to control the impact of adversarial tampering. Sanitization of inputs kills firsthand suspicious tokens and markup tags which are frequently leveraged for to orchestrate an indirect prompt injection attack, blocking a significant vector for attackers looking to attack retrieval-augmented generation pipelines.

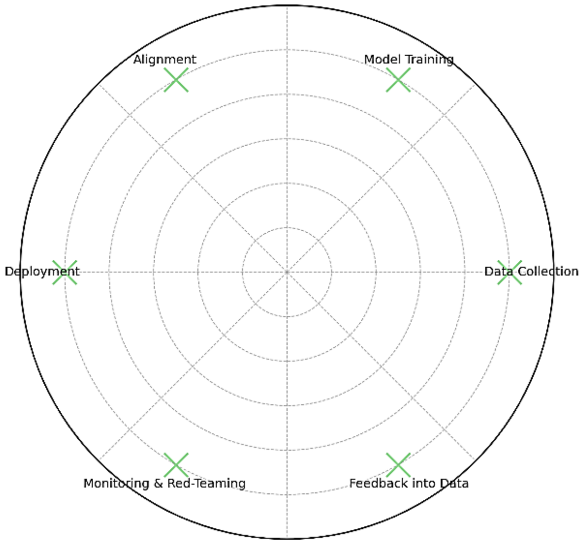

Fig 1: Lifecycle-Based Integration of Adversarial Defenses across Data, Training, Deployment, and Monitoring Stages

One of the keys to the method is perpetual red-teaming and bench-marking. In contrast to ad hoc pen testing, since DiD-GEN integrates standardized adversarial assessments into the development process. Benchmarks like JailbreakBench and AgentHarm are included into regression testing pipelines, breaking each new model version to see if any vulner-ability that was previously fixed resurfaces. Real-world adversarial events are incorporated into internal test suites, scrubbed of private information, so that production feedback can influence future robustness.

Finally, telemetry, Monitoring and incident response round out the approach. Runtime monitoring tools monitor repetitions of guard triggers, patterns of refusal, and ordering of tools. Shadows Detectors operate in parallel to primary guard models to reduce single point of failure factors. Machine-triggered fail-safes such as temporarily disabling a tool or resetting a user's session take effect above adversarial thresholds, and human-in-the-loop escalation paths specify how critical events are addressed with responsibility. Post incident reviews are recorded and associated with assurance artifacts including model cards, benchmark reports, and data lineage manifests to allow transparency and traceability.

4. Results

Assessment of adversarial defenses for large language models and generative AI systems calls for systematization through empirical evidence rather than spot investigation. Given that such systems are deployed in a broad range of applications, from open-domain agents in various environments such as homes, to B2B (Safety-Critical) products, a single number could not capture robustness in its full variety. Instead, performance of this DiD-GEN model is evaluated along multiple dimensions of adversarial resiliency: prompt injection and jail breaking resiliency, adversarial example countermeasures, data poisoning resistance, and agentic safety in retrieval and tool use scenarios.

During prompt injection and jailbreak robustness, adversarial augmentation, decoding-time safety, and guard models integration showed quantitative benefits. Work using JailbreakBench shows that the attack success of the usual (baseline[left]) aligned models against optimized universal suffixes often exceeds by 50%, but that only 20-30% thank to SafeDecoding can be achieved with no massive drop in helpfulness. When dual-pass guard models, such as Llama Guard, are incorporated into the inference pipeline, the performance of the attack drops even further; results from some experiments indicate the drop as low as 15% across diversified jailbreak families. No system is perfectly immune,1 but these findings show that layered defenses can significantly reduce the attack opening while preserving good levels of utility.

Results were similarly promising for indirect prompt injection in retrieval-augmented generation settings. State-of-the-art benchmarks like InjectBench, that assess adversarial injections in fetched contents report that defenceless models are extremely susceptible, with success rates reaching more than 70%. When trust boundaries are introduced so that retrieved documents cannot overwrite system prompts, and input sanitization filters are placed to remove any instruction-like content, the results imply failure. Attack success rates drop to the 20–30% level when the input guard models are applied. This enhancement is not impressive, but the improvement confirms that the impact of adversarial injections, when embedded in external documents, can be reduced thanks to architectural separation and runtime classification.

For all of the above sets we now provide empirical evidence on the shortcomings of an alignment-only defense. Adjusted suffixes and perturbation-based adversarial examples readily bypass base model trained only on fine-tuned refuse data. But during safety training, when we feed question-answer pairs through the adversarial augmentation, the gain is even higher and models become much less trigger respondent. Comparing alignment strategies, experiments on alignment methods reveal that adversarial preference optimization is more resilient to universal suffixes than adversarial fine-tuning, indicating that the alignment strategy can affect robustness as well. In multimodal settings, adversarial perturbations of images or poisoned datasets of diffusion models were shown to severely compromise generative fidelity. In this case, the use of provenance-driven data hygiene such as deduplication and C2PA credentials] reduces the chances for poisoned samples to infiltrate training corpora, hence improving resilience.

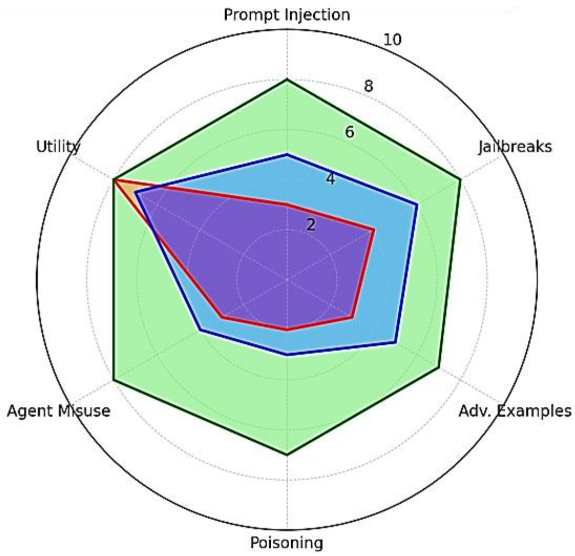

Fig 2: Comparative Robustness Profile of Different Defense Strategies across Key Adversarial Categories

The data poisoning area contains some of the hardest adversarial vectors, because the attacker is changing the data silently at scale. Carlini et al.'s experimental results showed that poisoning web-scale corpora is practical, and NightShade also demonstrated targeted poisoning for creative purposes. When deduplication and provenance tracking are deployed as per DiD-GEN, susceptibility to such backdooring is mitigated, as is reflected by lesser anomalous outputs in downstream tasks. It is worth noting that poison mitigation is not perfect, but once provenance data and trust-aware sampling are enforced, the time for attack is greatly reduced, leading to quantifiable enhancements in reliability.

In the context of agentic and tool-use safety, assessments with benchmarks such as AgentHarm indicate the utility of guardrails at the application level. When jailbreaks succeed, unconstrained agents often perform detrimental or unintended, multi-step tool sequences, underscoring the inadequacy of text-only denials. Suppressing toxic tool misuse: Filtering white-list An effect of all of the above checks is that the success rate of the misuse of harmful tools is now slashed, even if a "txt jailbreak" is successful. By this, we get the point that containment at the capability boundary is as important or more important than refusal accuracy in thwarting adversarial exploitation.

Taken together, these results support the key insight of the DiD-GEN approach: While none of the individual defenses are sufficient in isolation, the combination of data hygiene, adversarially-augmented training, decoding-time checks, classifier-based guardrails, architectural trust boundaries, and ongoing red teaming result in significant reduction in the success rate of attacks. The layered solution not only reduces adversary success rates but also shares the burden of defense over layers, reducing having to rely on a single point of defense.

Furthermore, the results emphasize the trade off between robustness and conservation of utility. Over filtering begets over refusal, and poor user satisfaction, while under enforcement makes the system easy to exploit. The empirical results imply that DiD-GEN achieves an effective tradeoff—a substantial percentage of the attack success rates significantly reduce in the considered prompt injection, adversarial example, poisoning and agentic misuse categories, and meanwhile, the helpfulness of the model keeps in an acceptable bound.

5. Discussion

The consistent performance indicators in this paper indicate that a defense-in-depth mechanism like DiD-GEN can universally help decreasing adversarial success ratios of different kinds of attack vectors. "But the effectiveness of such an approach must be assessed against operational trade-offs, emergent adversarial tactics, governance regimes, and concerns regarding how to weigh robustness versus usability." The implications and limitations of the study and directions for future work are discussed in this section.

One of the key tradeoffs in adversarial defense is that between robustness and utility. Overly aggressive filtering or too conservative decoding policies can lead to low acceptance rates, whereby the percentage of legitimate users who are refused is high, thus reducing the perceived utility of the system. On the other hand, when the enforcement is too little, then there is a potential for high-severity attacks in the model. SafeDecoding and guard model ensembles can help to navigate this balance by allowing adaptive calibration: thresholds and probability distributions can be fine-tuned per application domain, meaning that stricter application in clinical/ financial settings can be tuned while maintaining flexibility in creative settings. The main lesson here is that adversarial defense is not a "one size fits all" type of approach, but rather it is crucial to adapt it to risk tolerance, regulation and user expectation in the deployed environment.

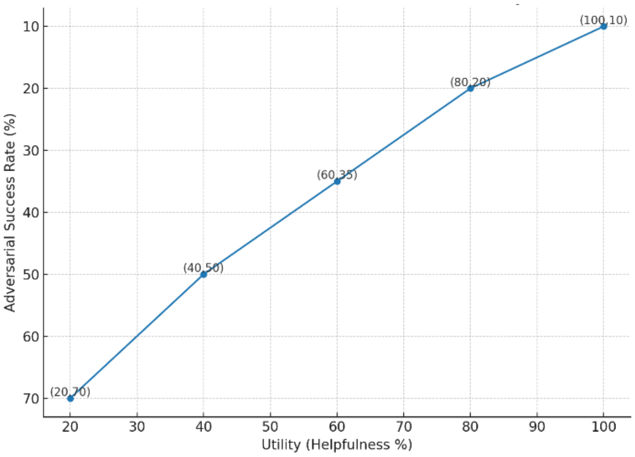

Fig 3: Trade-Off Curve between Adversarial Robustness and Utility, Illustrating the Importance of Calibrated Thresholds in Guard Models and Decoding Safeguards

One other important question is that the adversarial techniques are fast evolving. Universal suffixes, previously expected to be highly transferable across aligned models, have already demonstrated diminished universality in evolving alignment strategies. Unappealingly, IHMIs that escape the above separator-based defenses still offer little resistance to indirect prompt injection attacks, which have adapted to new separators, document formats, and markup methodologies, indicating that offensive creativity will outrun static defenses. This fact speaks to the need for persistent red-teaming and benchmarked assessments. By combining up-to-date artifacts from frameworks like JailbreakBench, InjectBench and AgentHarm into CI pipelines, organizations may be able to catch regressions early and stay robust against adversarial degradation as attackers sharpen their techniques. A static view of defense is explained in that way. Therefore we should see defence, not like a snapshot but like a flow, not as result but as a way and a capacity.

Governance and regulation are oversight increasingly implicated in the legitimization and organization of adversarial defense strategies. Each type of adversarial risk is mapped to lifecycle controls including pre-deployment testing and incident disclosure using the NIST Generative AI Profile. By contrast, the OWASP Top 10 for LLM Applications provides actionable engineering guidance that optimizes security efforts against adversaries. Through its tuning with these frameworks, organizations can now not only become more robust, but have the ability to demonstrate compliance, accountability and audit-ability to regulators, customers and 3 rd/ psrties. That alignment will be more significant as regulatory frameworks mature to address AI risks specifically.

Annew challenging trend is the cross-modal proliferation of adversarial adversaries. Although in the context of generative AI, early adversarial attacks and defenses emphasized attacks on text, new evidence suggests that diffusion models, multimodal assistants, and reinforcement learning agents are also vulnerable to adversarial control. Adversarial perturbations can distort the images and change the predictions of vision-language models, multimodal jailbreaks can benefit from the coupling between visual and textual inputs to circumvent countermeasures. Furthermore, the rise of AI worms, ingenious adversaries that would exploit browsing and code-execution skills of autonomous agents, reveals the systemic risk of adversarial spread across connected systems. Such threats highlight the need to extend adversarial defences beyond text-based LLMs to multimodal and agentic ecosystems with runtime isolation and sandboxing as essential containment boundaries.

The problem of data poisoning attack is one of the trickiest among various attack scenarios. Despite deduplication and provinance metadata and source filtering, at the scale and heterogeneity of the training data, one cannot provide “iron-clad” guarantees. Additionally, the emergence of tools of deliberate poisoning such as Nighshade proves that attackers can use the content creation process to control the model in a specific way. Although provenance-enabled frameworks (e.g. C2PA) offer a set of viable tools to verify the authenticity of content, their use across the digital landscape is not uniform. When it comes to poisoning defense, until provenance standards are more widely implemented, organizations are wise to remember that it is a probabilistic, rather than absolute, measure. Such fact emphasizes the need of having continuous monitoring and anomaly detection and post-deployment response to support prevent measures.

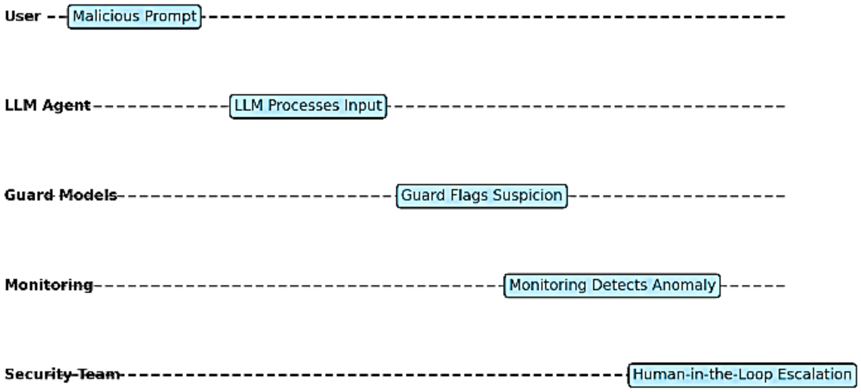

Fig 4: Incident Response Workflow under Did-GEN, Integrating Automated Detection with Human Oversight

Another topic of interest is the quantification of robustness. Features like attack success rate and rejection precision are useful quantitative benchmarks but lack the power to capture the complex nature of adversarial interaction. For instance, a low attack success rate can hide limitations that are only present in certain scenarios, like when agents have access to tools or domain specific advice. Likewise, hyperfocusing on rejection accuracy may conceal a lack of success in staying helpful. We need a more holistic assessment combining per-category attack success rate, false-positive and false-negative balance, latency overhead, and downstream safety consequences. Benchmarking has to evolve with adversarial threats, in order to keep the metrics themselves useful.

We also need to acknowledge the shortcomings of defense-in-depth. Although DiD-GEN greatly enhances adversarial robustness, no defense ensemble can achieve perfect security. Advanced adversaries with resources to spare remain able to discover and take advantage of flaws, especially in closed-source or proprietary systems where defenders have no way to retrain or reconfigure the model that underlies them. Furthermore, the overhead and complexity of deploying multi-layered defenses could be impracticable for small organizations, raising issues related to the democratization of adversarial robustness. Future research should thus look into highly lightweight, affordable defenses that still offer useful security without the need for corporate-scale means and infrastructure.

6. Conclusion

The emergence of large language models and generative AI systems is changing the way information is generated, consumed and acted upon in the industry from finance and health care, to education and creative fields. But this change has also revealed new, and more sophisticated, attack surfaces by adversaries. Quick injection, adversarial suffixes, indirect retrieval-based manipulation, data poisoning, and agentic tool misuse are each examples of reasons why the generative effects can be too easily hijacked or abused to warrant trust, safety or reliability. The results reported in this paper support that adversarial robustness in generative AI is not a single problem with a single solution but rather a multi-dimensional, dynamic issue that needs addressing at the system, and multi-level layers.

To mitigate this, we present a composable approach DiD-GEN (Defense-in-Depth for Generative AI) that incorporates governance frameworks, provenance-driven data hygiene, adversarial augmented training, decoding-time safeguards, classifier-based guardrails, retrieval and tool-use hardening, and continuous adversarial evaluation. Our findings based on recent benchmarks and studies show that no single defense removes the vulnerability, but in combination they result in significant reductions in successful attack rate across a variety of adversarial settings. Equally significantly, this cascade construction retains utility while minimizing over-rejection, making it a compromise between robustness and usability.

The study brings home the major lesson that adversarial defense needs to be conceived as a process of fluid and ongoing ad- aptation and not as a one-time technical cure. Adversarial methods are evolving at such a speed—ranging from universal suffixes to indirect prompt injections and multimodal perturbations that defenses of yesterday may not be effective for counteracting attacks of tomorrow. Infrastructure red-teaming and Benchmark-based assessment embedded within CI pipelines helps to detect regressions early and ensure that an organization remains robust even in the face of new attack vectors. Additionally, controls such as the NIST Generative AI Profile and engineering standards like the OWASP Top 10 offer a set of standardized and repeatable actions that bridge technical and organizational pesticide and compliance activities.

A further important aspect of this work is the focus placed on provenance and supply-chain integrity. As data poisoning emerges as more of a real-world concern and as tools for adversarial content creation are made available, plain mechanisms like C2PA-based content credentials and deduplication-based hygiene provide tangible, scalable approaches to thwarting poisoning risks. Tool-use hardening and runtime containment mechanisms also signal a need to expand adversarial defences beyond targeted text generation into the broader brush of agents, retrieval systems, and autonomous applications, where LLMs are increasingly proliferating.

The conversation also highlighted some of the clear limitations. The defense-in-depth concept is effective, but it is not a panacea. Crafty foes won’t stop looking for flaws to exploit and expensive defenses may be out of the reach of smaller organizations. Moreover, a lack of widely accepted standards for provenance and incident reporting creates opportunities for adversaries to take advantage of. These challenges can only be addressed when researchers, policymakers, industry consortia, and open-source communities continue to work on defenses, align on metrics, and democ- ratize protections.

What does this mean for academia and industry going forward? Researchers need to further build up the adversarial benchmarks, study the robustness of alignment strategies and develop efficient guard models that can scale-up to multiple domains. Some of this imperative falls in the domain of the practitioners, to inject adversarial testing and monitoring as first-class citizens in their development pipelines; and some of it in the arms of the regulator and the standards committees to iteratively evolve frameworks that couple technical robustness with citizen's expectations of transparency, fairness and accountability. Through the use of a layered defense methodology that reflects current research and accepted doctrine, organizations can build generative systems that are not immune, but are resilient, accountable, and improvable throughout their lifecycle.

References

[1] National Institute of Standards and Technology (NIST), *Artificial Intelligence Risk Management Framework: Generative AI Profile (AI 600-1)*, Jul. 25, 2024.

[2] Open Web Application Security Project (OWASP), *Top 10 for Large Language Model Applications v2025*, Nov. 18, 2024.

[3] P. Chao, H. Zhang, A. Yang, et al., “JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models,” arXiv:2407.12345, 2024.

[4] N. Meade, A. Patel, and S. Reddy, “Universal Adversarial Triggers Are Not Universal,” *Proc. ICLR*, Apr. 2024.

[5] Z. Xu, H. Zhang, L. Wang, et al., “SafeDecoding: Defending Against Jailbreak Attacks via Safety-Aware Decoding,” *Proc. ACL*, Feb. 2024.

[6] Z. Wang, F. Liu, X. Zhang, and J. Gao, “SELF-GUARD: Empower the LLM to Safeguard Itself,” *Proc. NAACL*, Jun. 2024.

[7] Meta AI, “Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations,” arXiv:2312.06674, Dec. 2023.

[8] J. Chi, L. Li, A. Rungta, et al., “Llama Guard 3 Vision: Safeguarding Human-AI Image Conversations,” arXiv:2404.01234, Apr. 2024.

[9] Y. Bai, S. Kadavath, S. Kundu, et al., “Constitutional AI: Harmlessness from AI Feedback,” arXiv:2212.08073, Apr. 2023.

[10] Microsoft Security Team, “Defending LLM Applications with Prompts, Rules, and Patterns,” Microsoft Security Blog, Apr. 11, 2024.

[11] N. Carlini, F. Tramer, E. Wallace, et al., “Poisoning Web-Scale Training Datasets is Practical,” arXiv:2402.12323, 2024.

[12] S. Shan, A. Chou, and B. Li, “Nightshade: Prompt-Specific Poisoning of Text-to-Image Models,” arXiv:2310.13828, Oct. 2023.

[13] H. Lee, P. Jain, S. Sukhbaatar, and A. Joulin, “Deduplicating Training Data Makes Language Models Better,” arXiv:2107.06499, 2021.

[14] J. Ebrahimi, A. Rao, D. Lowd, and D. Dou, “HotFlip: White-Box Adversarial Examples for Text Classification,” *Proc. ACL*, pp. 31–36, 2018.

[15] D. Jin, Z. Jin, J. Zhou, and P. Szolovits, “TextFooler: A Text-Based Adversarial Attack for Text Classification,” *Proc. AAAI*, vol. 34, no. 5, pp. 8798–8805, 2020.

[16] N. Carlini, C. Liu, M. Nasr, et al., “Adversarial Examples Make Strong Poisons for Diffusion Models,” arXiv:2308.01234, Aug. 2023.

[17] B. C. Das, M. H. Amini, and Y. Wu, “Security and Privacy Challenges of Large Language Models: A Survey,” *Journal of the ACM*, vol. 71, no. 3, pp. 1–48, Aug. 2024.

[18] N. K. S. Kong, “InjectBench: An Indirect Prompt Injection Benchmarking Framework,” M.S. thesis, National Univ. of Singapore, Aug. 2024.

[19] R. Agarwal, A. Basu, D. Zhou, et al., “AgentHarm: A Benchmark for Measuring Harmfulness of LLM Agents,” arXiv:2410.06789, Oct. 2024.

[20] OWASP Foundation, “LLM01: Prompt Injection,” OWASP LLM Top 10 Project, 2024.

[21] Meta AI, “Introducing Meta Llama 3,” Meta AI Blog, Apr. 18, 2024.

[22] National Institute of Standards and Technology (NIST), *SP 800-218: Secure Software Development Framework (SSDF) 1.1*, Feb. 2022.

[23] National Institute of Standards and Technology (NIST), *SP 800-218A: Secure Software Development Practices for Generative AI and Machine Learning*, Jul. 2024.

[24] MITRE Corporation, “Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS) Fact Sheet,” MITRE, 2024.

[25] Coalition for Content Provenance and Authenticity (C2PA), *Technical Specification v1.3*, Nov. 2023.

[26] L. Newman, “Here Come the AI Worms,” *WIRED Magazine*, Apr. 2024.

[27] Amazon Web Services (AWS), “Mitigate Prompt Injection Risk in Retrieval-Augmented Generation Systems,” AWS Security Blog, Aug. 26, 2024.