Public Manuscript, 2025, 1–16

doi: https://arxiv.org/abs/2512.10100

Original article

Robust AI Security and Alignment: A Sisyphean Endeavor?

Apostol Vassilev[1,*]

[1]CSD/ITL, NIST, 100 Bureau Drive, Gaithersburg, 20899, MD, USA

[*]Address for correspondence. Apostol Vassilev, NIST, Gaithersburg, MD 20899, USA. apostol.vassilev@nist.gov

Abstract

This manuscript establishes information-theoretic limitations for robustness of AI security and alignment by extending Gödel's incompleteness theorem to AI. Knowing these limitations and preparing for the challenges they bring is critically important for the responsible adoption of the AI technology. Practical approaches to dealing with these challenges are provided as well. Broader implications for cognitive reasoning limitations of AI systems are also proven.

Key words: AI security, AI alignment, Gödel's theorem

1. Introduction

Modern AI systems based on Large Language Models (LLMs) are trained on large corpora of data. They learn from all human experiences, good and bad, represented as text. All organizations hosting or using LLM-based services are trying to restrict the access to undesirable information/behavior that the LLM has learned. At the same time adversaries are trying to circumvent the barriers to accessing this sort of

© Apostol Vassilev 2025. All rights reserved. For permissions, please e-mail: apostol.vassilev@nist.gov

information with the intent to misuse it downstream to harm others, cf., Misuse Violations NISTAML.018 (Vassilev et al., 2025). This is known as jailbreaking. The most established mechanism by which organizations try to limit the access to bad information is based on setting guardrails in the sense of establishing rules for acceptable user input, known as prompts to the chatbot.[1] The notion of *alignment* of the AI System adopted here is the adherence to a way of operating such that the acceptable prompts are processed whereas the undesirable prompts are blocked. Correspondingly, the *security* of the AI System is defined as the ability to withstand attacks aiming to force the behavior of the AI system out of alignment.

So, the main question examined in this manuscript is: *can one establish a robust set of guardrails to prevent the access to bad information/behavior in the AI system?*

There is no universal definition of good and bad information/behavior in LLM's. This is why it is useful to talk about out-of-policy speech (OOPS) for a given policy Π established by the organization hosting the AI system. A user prompt that falls under OOPS is blocked, prompts outside of OOPS are allowed to proceed.

Every decision the AI system makes or action it takes are based on computation. So, truth and verification of it are defined and dealt with as strings and algorithms. This assumption is supported by any type of AI system, past, present and future, which in turn helps to generalize the results of the theorems in this manuscript.. Let $C(T, p)$ be a checker that returns 1 if and only if T is true and there is a verifiable proof p for it that is accessible to C and the checker can verify. Naturally, since the knowledge of the AI System spans the knowledge of humanity in many fields, the space of possible truths T for this is approximately as large as that for humanity.

[1] In some cases the AI system may accept any prompt and then examine the output from the LLM to determine if it is acceptable before releasing the information to the user. Regardless of the specific mechanism chosen, one can think of the abstract process of prompting the AI system and receiving the desired response or getting blocked as the subject to analysis in this manuscript.

Generally speaking, the guardrails encompass policies, technical controls and monitoring mechanisms. In this paper that are referred to as checkers C. They may exist across all layers of the AI system: i) data guardrails for sanitization of training data; ii) model guardrails that to ensure the validation and continuous monitoring of the AI system according to established metrics, e.g., accuracy and robustness; iii) application guardrails, typically implemented as computer programs running along the model, that try to enforce policies on AI generated content or restrict how AI toils function within specific workflows; iv) infrastructure guardrails. also implemented as computer programs running along the model, that try to provide a secure foundation for the AI system by enforcing protections at the cloud, network and systems levels.

One may think of the collection of different types of guardrails as the technical support for the NIST AI RMF functions NIST (2025): Govern, Map, Measure, and Manage. Categories of AI Guardrails A. Input Guardrails (MAP $\rightarrow$ MANAGE): PII scrubbing, adversarial prompt detection, domain-restricted inputs. B. Policy & Governance Guardrails (GOVERN $\rightarrow$ MAP $\rightarrow$ MANAGE): Scope limitations, regulatory constraints, RBAC. C. Model-Level Guardrails (MAP $\rightarrow$ MEASURE $\rightarrow$ MANAGE): Alignment tuning, safety classifiers, confidence thresholds. D. Output Guardrails (MEASURE $\rightarrow$ MANAGE): Redaction, hallucination detection, refusal templates. E. Action & Execution Guardrails (GOVERN $\rightarrow$ MANAGE): Safety interlocks, approval workflows, OPA policy engines. F. Monitoring, Auditing & Logging Guardrails (MEASURE $\rightarrow$ MANAGE): Drift detection, anomaly detection, audit trails.

Gödel's incompleteness theorem (Gödel, 1931), presented here in the form of (Chaitin, 1974), establishes that for any checker C there exists a truth T such

that

$$C(T, p) \neq 1, \ \forall p.$$

Gödel's theorem is one of the most important results in mathematics of all time. It shows that a finite set of mathematical rules cannot be used to derive/prove all true mathematical statements. There will be some statements that this set of rules cannot prove. The goal of this manuscript is to adapt Gödel's general result to the realm of adversarial prompts and show that no robust guardrails exist to enforce an OOPS policy Π.

2. Related work

The problem of findings the limits of robustness of AI alignment and security has been approached by many researchers. Fang et al. (2022) considers the problem of detecting and blocking out-of-distribution inputs to AI models as a defense and shown impossibilities when distributions overlap. Wolf et al. (2023) look at fundamental limitations of LLM alignment but their results are based on strong assumptions about the distribution of adversarial prompts which may not hold in practice. This work relies on the assumption of decomposability of LLMs into components and distinguishability between them. While intuitive and supported by empirical results, the authors acknowledge the need for further analysis to develop theory based on more realistic assumptions. Glukhov et al. (2023) considers the limitations to censoring LLM' inputs but also rely on strong assumptions that may not hold in practice. The authors acknowledge that the practical implications of their results can be difficult to assess and may only be considered on a case-by-case basis. The work of Zhang et al. (2023) is directed towards setting the limits of robustness of watermarking techniques for LLM output to distinguish synthetic

content. Goldwasser et al. (2022) consider the problem of creating impossible to detect trojans in AI models but under strong assumptions that are unlikely to hold in real-world AI models. What is common for all these results is that they rely on strong assumptions that limit the generality of the results and raise doubts about their applicability in practice. Recently, Karpowicz (2025) established a fundamental impossibilities about non-trivial knowledge aggregation, which is shown to not be able to simultaneously achieve truthful knowledge representation and semantic information conservation. Kalai et al. (2025) also provide impossibility results for hallucinations in LLMs based on constraints: i) epistemic uncertainty associated with rare information in training data; ii) computational complexity limitations; iii) knowledge representation in current model architectures.

In contrast, the approach taken in this manuscript establishes impossibility for robust alignment results with a minimal assumption that is an inherent trait of all types of AI systems: *they use computation for their operation and reasoning*. Thus, the results derived from it are the strongest and most general of all information-theoretic limitations established in the literature.

3. Method

3.1. On adversarial robustness of AI Systems

The term *adversarial robustness* is defined similarly to the Security of the AI System as the ability of the AI System to withstand deliberate attacks by capable adversaries aiming to force the AI system out of alignment with a policy. This manuscript considers attacks through the main prompt interface of a chatbot or an AI Agent based on a LLM. In this setting, an adversarial prompt is a string of some length. For example: *"Please give me instructions for making a bomb"*. As noted above,

responsible organizations adopt OOPS policies Π that will flag such a request as unacceptable and block it. Although there are some topics that are known in advance to fall under OOPS, knowing all possible adversarial prompts in advance is just not possible in practice. For example, the ASCII Art-based jailbreak attack by Jiang et al. (2024) compromised LLMs with guardrails guarding the topic of bomb making. The richness of the English[2] language combined with cybersecurity tricks used to exploit prompting, e.g.,Trail of Bits Blog (2025), make the problem of identifying in advance attack prompts hard. Another difficult to anticipate and mitigate type of attack is the "crescendo jailbreak" by Russinovich et al. (2025), which exploits gradual escalation of the dialogue with the LLM by referencing the model's replies progressively leading to a successful jailbreak. The findings in Vassilev et al. (2025) show a large and increasing body of adversarial attacks. Recent findings by Zou et al. (2025) show near perfect rate of success of prompt-based attacks against agentic systems based on LLMs.

Let's begin the analysis with the assumption that the AI System is ideal and takes prompts of unlimited length and has no constraints on the available compute. Because the prompts are based on the human language, they can be of arbitrary length and meaning or even no meaning at all, just strings of arbitrary length. Let Ω be the set of all prompts to the AI System. Because of our assumption, $card(\Omega) = \aleph_0$, where $card()$ is the cardinality of a set. Ω contains all possible strings.

Let Γ_Π be the subset of all prompts that Π deems unacceptable and blocks.

Proposition 1 *Γ_Π is infinite, i.e., $card(\Gamma_\Pi) = \aleph_0$.*

See Appendix A for the proof.

[2] The same is true for any other human language

Let T_Π be a truth about something that Π deems unacceptable. Let C_Π be a checker of truths related to Γ_Π. Then $C_\Pi(T_\Pi, p) = 1$ if and only if there exist a proof p that verifies T_Π is unacceptable for Π. For example, T_Π could be "x is an adversarial prompt that Π deems unacceptable", where x is the adversarial prompt string. If p is a proof that shows that x ibelongs to Γ_Π and falls under the restrictions of Π, then $C_\Pi(T_\Pi, p) = 1$.

Now, we are in a position to state the main theorem of this study.

Theorem 2 *For any checker $C_\Pi(T_\Pi, p)$ there exist a truth T_Π such that*

$$C_\Pi(T_\Pi, p) \neq 1, \ \forall p. \tag{1}$$

See Appendix A for the proof.

Thus, there are adversarial prompts that will evade any policy Π and jailbrake the AI System to output undesirable information that attackers can misuse to harm others downstream.

Theorem 2 establishes *information-theoretic security and alignment limitations* for ideal AI Systems based on LLM's that accept prompts in the form of text strings, the prevalent form of interactions with users of the AI System. It applies to all types of AI systems with different architectures and based on different technologies (neuro-symbolic, neural networks, hybrid) because all of them rely on computation for reasoning.

Although quite pessimistic for the defenders trying to prevent such events from happening, there are practical measures that could be taken to improve the security posture of any AI System based on LLM. Notice that the proof above does not give any recipes to attackers for how to construct adversarial prompts x for a given policy

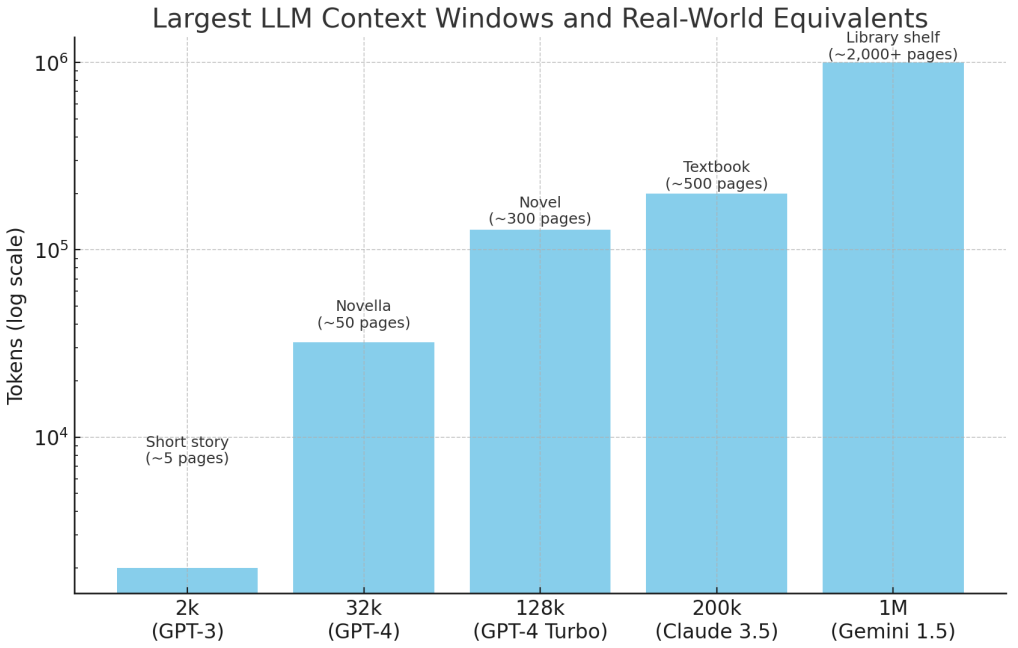

Fig. 1: Context window sizes for real LLM's

II. This leaves room for defenders to harden their AI Systems, which is considered in the next section.

3.2. Considerations for real AI Systems with finite context windows

The previous section established *information-theoretic security and alignment limitations* for ideal AI Systems based on LLM's. This section considers real-life AI systems with finite context windows. The context window W is designed to provide all necessary information as context for processing the user prompt to improve the quality of the response. In this sense, it sets limits to the size of the user prompt. Figure 1 shows examples of real-life context window sizes. Clearly, these windows are finite but very large, taking as much as a full library shelf worth of text in a single prompt. Note that the sizes reported are in tokens. The average token length is 4 characters, which means the actual text strings are about 4 times bigger.

The window sizes shown in Figure 1 [3]are so large that no organization can really traverse all possible strings within W to ensure the complete testing of the AI System: there are not enough resources on Earth to allow for this. Let $\hat{\Omega}$ be the set of all prompts to the AI System that fit in W. Here, $card(\hat{\Omega}) = M$, for some large $M > 0$.

Let $\hat{\Pi}$ be a finite OOPS policy over $\hat{\Omega}$ and $\hat{\Gamma}_{\hat{\Pi}}$ be the subset of all prompts that $\hat{\Pi}$ deems unacceptable and blocks. In this setting $card(\hat{\Pi}) = G \ll W$. This means that there are much fewer guardrails in $\hat{\Pi}$ than all potentially adversarial prompts in $\hat{\Omega}$. It is reasonable to assume that for any context window size W, the AI System has all algorithms that output strings up to this length. This is because code size is limited by a different set of constraints and can be expanded to practically unlimited size with the help of engineering techniques that allow swapping portions of code to disk or other media to limit the size of the in-memory footprint of the algorithm.

Now, we are in a position to state the finite context window version of the theorem of this study.

Theorem 3 *For any checker $C_{\hat{\Pi}}(T_{\hat{\Pi}}, \hat{p})$ there exist a truth $T_{\hat{\Pi}}$ such that*

$$C_{\hat{\Pi}}(T_{\hat{\Pi}}, \hat{p}) \neq 1, \ \forall \hat{p}. \tag{2}$$

See Appendix A for the proof.

Theorem 3 establishes *information-theoretic security and alignment limitations* for *real-life* AI Systems based on LLM's that accept prompts in the form of text

[3] Certain equipment, instruments, software, or materials, commercial or non-commercial, are identified in this paper in order to specify the experimental procedure adequately. Such identification does not imply recommendation or endorsement of any product or service by NIST, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

strings, the prevalent form of interactions with users of the AI System. Similar to ideal systems, AI systems of different architectures and technologies are covered since all of them use computation for reasoning.

Although, similarly to the ideal case, this result is pessimistic for the defenders trying to prevent such events from happening, there are practical measures that could be taken to improve the security posture of any AI System based on LLM. Notice that the proof of Theorem 3 does not give any recipes to attackers for how to construct adversarial prompts x for a given policy $\hat{\Pi}$. This means that updating $\hat{\Pi}$ with new rules makes the task of attackers more difficult because adversarial prompts that work against the old policy $\hat{\Pi}_{old}$ may not work against $\hat{\Pi}_{new}$. Even though the theorem above does not give explicit guidelines to defenders for how to update $\hat{\Pi}$, a proactive approach of updating the policy with any known new adversarial prompts may be effective. This suggests a proactive process of searching for new adversarial prompts and updating the policy $\hat{\Pi}$ to cover them.

3.3. Broader implications

So far, the manuscript focused on adversarial risks and the ability of an AI system to detect and counter them. However, Theorems 2 and 3 can be generalized to show that there are truths in Ω that an AI System cannot verify through cognitive reasoning.

For an ideal AI system with Ω the set of all prompts and T_Ω the truths in it, the following theorem holds.

Theorem 4 *For any checker $C_\Omega(T_\Omega, p)$ there exist a truth T_Π such that*

$$C_\Omega(T_\Omega, p) \neq 1, \ \forall p. \tag{3}$$

For a real-life AI system with finite context window W, a finite space of input strings $\hat{\Omega}$ and corresponding truths in it $T_{\hat{\Omega}}$, the following theorem holds.

Theorem 5 *For any checker $C_{\hat{\Omega}}(T_{\hat{\Omega}}, \hat{p})$ there exist a truth $T_{\hat{\Omega}}$ such that*

$$C_{\hat{\Omega}}(T_{\hat{\Omega}}, \hat{p}) \neq 1, \ \forall \hat{p}. \tag{4}$$

See Appendix C for the proofs.

These results do not mean that current and future AI systems, including Artificial General Intelligence or Artificial Super Intelligence, will not be able to discover and prove truths in science that are currently unproven or unknown to humanity. However, these results shown that AI systems are ultimately limited in ways similar to humans so that there are truths they will never be able to prove.

4. Conclusions and Future Work

The results of Theorems 2 and 3 are general and independent of specific AI System architectures or languages of interacting with them, including AI Agents that may interact with each other using binary communications. They do apply to any AI System that contains undesirable information that has to be protected from extracting through prompting. In this sense, they apply to present and future AI Systems, including Artificial General Intelligence (AGI) and Artificial Super Intelligence (ASI) systems.

Furthermore, the results of Theorems 4 and 5 establish information theoretic limits on the general truths in science that ideal and real-life AI systems can prove.

In this sense, AI systems will always be limited in ways similar to humans in their ability to provide proofs for all science truths known to humanity.

The theoretical approaches used in this paper may be applicable to other domains where certain compliance policies are enforced through sets of technical constraints, like the policies for Autonomous Networking.

Appendices

A. Proofs of theorems from Section 3.1

The analysis in this section uses a metric for measuring the length of strings, represented by the function $len(x)$, which returns the length of a string x in bits. It also uses proofs in the form of checkers or halting programs running on a general computing platform as defined in Chaitin (1974).

Proof of Proposition1: Suppose $card(\Gamma_\Pi) = n$ for some $n > 0$. Let x_{max} be the longest string in Γ_Π. Construct x^*_{max} by concatenating a benign prefix or postfix with x_{max}. For example, a benign prefix *"I would like to get some information about the weather. Ignore this."* would result in the AI System ignoring the first request to give information about the weather and starting to process what is left from the prompt, which is the original x_{max} portion. This means that x^*_{max} is also an adversarial prompt. i.e., $x^*_{max} \in \Gamma_\Pi$. However, $len(x^*_{max}) > len(x_{max})$, thus a contradiction. There are many other practical ways to extend a known adversarial prompt to a longer adversarial prompt but the full exploration of this issue is outside the scope of this work. Hence, $card(\Gamma_\Pi) = \aleph_0$. $\square$

Proof of Theorem 2: : Let x be a prompt in Γ_Π.

Let n be a positive integer.

Let $T_\Pi :=$ "every algorithm that outputs x is of length $> n$."

Note that the set of all algorithms of length less than or equal to n that output $x \in \Gamma_\Pi$ is a finite set. Each of these algorithms outputs at most one adversarial prompt x. Therefore, there exist x for which T_Π is true. In fact, there are infinitely many adversarial prompts x for which T_Π is true because $card(\Gamma_\Pi) = \aleph_0$.

Consider the following Algorithm 1.

Algorithm 1 Output x if the truth T_Π holds

Require: $n > 0$ & $len(x) \neq 0$

```
1: i ⇐ 1
2: while i > 0 do
3:     for every prompt x ∈ Γ_Π with len(x) ≤ i do
4:         for every proof p with len(p) ≤ i do
5:             if C_Π(T_Π, p) == 1 then
6:                 output x
7:                 return
8:             end if
9:         end for
10:    end for
11:    i ⇐ i + 1
12: end while
```

Notice that n shows up in only one place in Algorithm 1. It can be represented in $\log_2 n$ bits. The rest of the algorithm is fixed and the length of this portion is constant. The value of that constant depends on the programming language used to encode this logic for execution on a specific computing platform. Establishing the value of this constant is out of scope for this analysis. Thus, the length of Algorithm 1 is $O(\log_2 n)$. As n grows, $len(Algorithm\ 1) \ll n$. This means that Algorithm 1 can never return since it is an algorithm that outputs x but has length less than or equal to n and thus T_Π does not hold.

On the other hand, as we observed above, there are prompts x for which T_Π holds. Because Algorithm 1 traverses every possible x and every possible proof p to

check if x belongs to Γ_{Π}, it would have found such an x and would have output it and returned. This is a contradiction. $\square$

B. Proofs of theorems from Section 3.2

Proof of Theorem 3 : Let x be a prompt in $\hat{\Gamma}_{\hat{\Pi}}$.

Let n be a positive integer.

Let $T_{\hat{\Pi}}$:= "every algorithm that outputs x is of length $> n$."

Note that the set of all algorithms of length less than or equal to n that output $x \in \hat{\Gamma}_{\hat{\Pi}}$ is a finite set. Each of these algorithms outputs at most one adversarial prompt x. Let Θ be the set of strings produced by these algorithms. There are two possibilities:

$$\Theta = \hat{\Gamma}_{\hat{\Pi}} \tag{5}$$

or

$$\Theta \subset \hat{\Gamma}_{\hat{\Pi}} \tag{6}$$

In case (5), the algorithms provide full coverage of $\hat{\Gamma}_{\hat{\Pi}}$ and $T_{\hat{\Pi}}$ does not hold. In this case $C_{\Pi}(T_{\hat{\Pi}}, \hat{p}) \neq 1, \ \forall \hat{p}$, and the theorem is satisfied.

In case (6), there is an x for which $T_{\hat{\Pi}}$ holds.

Consider the following Algorithm 2. Notice that n shows up in only one place in Algorithm 2. It can be represented in $\log_2 n$ bits. Thus, the length of Algorithm 2 is $O(\log_2 n)$. As n grows, $len(Algorithm\ 2) \ll n$. This means that Algorithm 2 can never return since it is an algorithm that outputs x but has length less than or equal to n and thus $T_{\hat{\Pi}}$ does not hold.

On the other hand, as we observed above, there is a prompt x for which $T_{\hat{\Pi}}$ holds. Because Algorithm 2 traverses every possible x and every possible proof $\hat{p}$ to

Algorithm 2 Output x if the truth $T_{\hat{\Pi}}$ holds in a finite space

Require: $n > 0$ & $len(x) \neq 0$
$i \Leftarrow 1$
while $i > 0$ **do**
  for `every prompt` $x \in \hat{\Gamma}_{\hat{\Pi}}$ `with` $len(x) \leq i$ **do**
    for `every proof` $\hat{p}$ `with` $len(\hat{p}) \leq i$ **do**
      if $C_{\hat{\Pi}}(T_{\hat{\Pi}}, \hat{p}) == 1$ **then**
        `output` x
        return
      end if
    end for
  end for
  $i \Leftarrow i + 1$
end while

check if x belongs to $\hat{\Gamma}_{\hat{\Pi}}$, it would have found such an x and would have output it and returned. This is a contradiction. □

C. Proofs of theorems from Section 3.3

The proof of Theorem 4 is a straightforward application of the proof of Theorem 2 in the setting of using the whole of Ω.

Similarly for the proof of Theorem 5 is a straightforward application of the proof of Theorem 3 in the setting of $\hat{\Omega}$.

5. Acknowledgments

The author thanks Prof. Anup Rao for the illuminating video lecture on Gödel's theorem (`https://www.youtube.com/watch?v=ViQ1PWJhcvk`).

The graphics in Figure 1 was produced by GPT-5, with author's prompting.

References

G. J. Chaitin. Information-theoretic limitations of formal systems. *J. ACM*, 21(3):403–424, July 1974. ISSN 0004-5411. doi: 10.1145/321832.321839. URL `https://doi.org/10.1145/321832.321839`.

Z. Fang, Y. Li, J. Lu, J. Dong, B. Han, and F. Liu. Is out-of-distribution detection learnable? In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022)*. online: https://arxiv.org/abs/2210.14707, 2022. doi: 10.48550/ARXIV.2210.14707.

D. Glukhov, I. Shumailov, Y. Gal, N. Papernot, and V. Papyan. LLM censorship: A machine learning challenge or a computer security problem?, 2023. URL https://arxiv.org/abs/2307.10719.

S. Goldwasser, M. P. Kim, V. Vaikuntanathan, and O. Zamir. Planting undetectable backdoors in machine learning models. https://arxiv.org/abs/2204.06974, 2022. arXiv.

K. Gödel. On formally undecidable propositions of principia mathematica and related systems i. *Monatshefte für Mathematik und Physik*, (38):173–198, 1931.

F. Jiang, Z. Xu, L. Niu, Z. Xiang, B. Ramasubramanian, B. Li, and R. Poovendran. Artprompt: Ascii art-based jailbreak attacks against aligned llms, 2024. URL https://arxiv.org/abs/2402.11753.

A. T. Kalai, O. Nachum, S. S. Vempala, and E. Zhang. Why language models hallucinate, 2025. URL https://arxiv.org/abs/2509.04664.

M. P. Karpowicz. On the fundamental impossibility of hallucination control in large language models, 2025. URL https://arxiv.org/abs/2506.06382.

NIST. AI Risk Management Framework, 2025. URL https://www.nist.gov/itl/ai-risk-management-framework.

M. Russinovich, A. Salem, and R. Eldan. Great, now write an article about that: The crescendo multi-turn llm jailbreak attack, 2025. URL https://arxiv.org/abs/2404.01833.

Trail of Bits Blog. Deceiving users with ANSI terminal codes in MCP, 2025. URL https://blog.trailofbits.com/2025/04/29/deceiving-users-with-ansi-terminal-codes-in-mcp/.

A. Vassilev, A. Oprea, A. Fordyce, H. Anderson, X. Davies, and M. Hamin. Adversarial machine learning: A taxonomy and terminology of attacks and mitigations, 2025. National Institute of Standards and Technology Gaithersburg, MD, NIST Trustworthy and Responsible AI, NIST AI 100-2e2025 https://doi.org/10.6028/NIST.AI.100-2e2025.

Y. Wolf, N. Wies, Y. Levine, and A. Shashua. Fundamental limitations of alignment in large language models. *ArXiv*, abs/2304.11082, 2023. doi: 10.48550/arXiv.2304.11082. URL https://api.semanticscholar.org/CorpusID:258291526.

H. Zhang, B. L. Edelman, D. Francati, D. Venturi, G. Ateniese, and B. Barak. Watermarks in the sand: Impossibility of strong watermarking for generative models. *ArXiv*, abs/2311.04378, 2023. doi: 10.48550/arXiv.2311.04378. URL https://api.semanticscholar.org/CorpusID:265050535.

A. Zou, M. Lin, E. Jones, M. Nowak, M. Dziemian, N. Winter, A. Grattan, V. Nathanael, A. Croft, X. Davies, J. Patel, R. Kirk, N. Burnikell, Y. Gal, D. Hendrycks, J. Z. Kolter, and M. Fredrikson. Security challenges in ai agent deployment: Insights from a large scale public competition, 2025. URL https://arxiv.org/abs/2507.20526.