

Defending Language Models: Safeguarding Against Hallucinations, Adversarial Attacks, and Privacy Concerns

Yetian He ^a

School of the Gifted Youth, University of Science and Technology of China, Anhui, China

Keywords: Large Language Model, Security, Hallucination, Adversarial Attacks, AI Alignment.


Abstract: While Large Language Models (LLMs) have garnered immense popularity, they also bring forth significant safety concerns. If LLMs are not disseminated to users in a secure and reliable manner, their continued development and widespread adoption could encounter substantial opposition and impediments. Hence, the primary aim of this survey is to systematically organize and consolidate current studies on LLM security to facilitate further exploration in this critical area. The article meticulously examines various security issues associated with LLMs, categorizing them into distinct sub-problems. It delves into the phenomenon of LLM hallucinations, elucidates mitigation strategies, explores adversarial attacks targeting language models, and evaluates defence mechanisms against such attacks. Furthermore, it discusses research pertaining to Artificial Intelligence (AI) alignment and security concerns in the context of LLMs. Additionally, the survey presents findings from relevant experiments to underscore the significance of addressing LLM security. By providing a comprehensive overview of LLM security, this paper aims to expedite researchers' understanding of this burgeoning field and catalyse advancements in ensuring the secure deployment and utilization of LLMs.

1 INTRODUCTION

A Large Language Model (LLM) is an Artificial Intelligence (AI) model that use natural language processing (NLP) technology to comprehend human language and provide replies, therefore challenging humans to envision engaging with AI. LLMs have emerged as a powerful tool for various applications, including text generation, translation, and code completion. These capabilities are developed by LLMs through a computationally expensive process of self-supervised and semi-supervised training, whereby they acquire statistical correlations from textual sources (Radford, 2019). Among the numerous LLMs, Chat Generative Pre-trained Transformer (ChatGPT) stands out as the most renowned and remarkable one. The fundamental concept behind GPT models is to condense global knowledge into the decoder-only Transformer model by language modelling. This allows the model to retain and recall the meaning of global knowledge, enabling it to function as a versatile solution for many tasks (Zhao, 2023). With a larger scale of parameters, GPT-3 utilizes in-context learning (ICL) for

predicting the appropriate task resolution and a three-stage reinforcement learning from human feedback (RLHF) algorithm (Ouyang, 2022) for improvement of human alignment. In March 2023, GPT-4 (Achiam, 2023), a large multimodal model, marking another significant advancement, demonstrates a level of performance comparable to that of humans on a range of professional and academic assessments. However, the rapid rate at which development teams are using LLMs has surpassed the implementation of thorough security standards, resulting in numerous applications being exposed to significant security risks. Research into patching vulnerabilities and defending against attacks has become more urgent.

In October 2023, Open Web Application Security Project (OWASP) released its top 10 LLM application vulnerabilities: 1. Prompt Injection; 2. Insecure Output Handling; 3. Training Data Poisoning; 4. Model Denial of Service; 5. Supply Chain Vulnerabilities; 6. Insecure Plugin Design; 7. Sensitive Information Disclosure; 8. Excessive Agency; 9. Overreliance; 10. Model Theft. This essay will not provide a detailed analysis of all 10 vulnerabilities, but instead will concentrate on

^a <https://orcid.org/0009-0001-3561-6767>

selected AI inherent ones. The security issues posed by the flaws in LLMs themselves are clearly novel and worth investigating. For instance, Hallucination in LLMs emphasizes the critical issue that content generated by LLMs that is inconsistent with real-world facts or user inputs, presenting significant obstacles for the practical implementation and raises questions regarding the dependability of LLMs in real-life situations (Agrawal, 2023). Additionally, adversarial attacks pose a threat that arises as a result of the vulnerabilities in LLMs, by identifying malicious inputs that cause LLMs to produce undesirable outputs (Zou, 2023). Jailbreaking, a typical example of adversarial attacks, capitalizes on the conflict between a model’s capabilities and safety goals or mismatched generalization of LLMs to output harmful content (Wei, 2024). More vulnerabilities and attacks will be described later.

Currently, there are several defence systems in place to safeguard LLMs from security issues. In response to the problem of hallucinations, there exist multiple techniques for identifying hallucinations in order to proactively mitigate their occurrence (Varshney, 2023). Improving data quality, model architecture, decoding strategies and enhancing contextual attention are regarded as an effective way to mitigate hallucinations. To address different adversarial attacks, the most prevalent approach to reduce the risks of these assaults is to train the model with samples specifically designed to simulate attacks, which is referred to as adversarial

training, but resulting in a compromise between the model’s performance and its resilience (Jain, 2023).

This survey builds upon notable preliminary efforts, striving to offer a comprehensive overview of the vulnerabilities inherent in LLMs and the strategies employed to address them. Focusing on current research trends, it identifies hallucinations and adversarial attacks as the primary concerns in LLM security. As show in Figure 1, the paper categorizes LLM vulnerabilities into misinformation, adversarial attacks, and other risks, examining corresponding defense mechanisms. The structure and content of the paper are outlined as follows: Section 2 elaborates on factuality and faithfulness hallucinations, along with preventative measures. Section 3 explores five specific forms of adversarial attacks and their defenses. Section 4 addresses AI alignment and privacy as critical LLM security topics. Section 5 summarizes experiment outcomes and conducts a comparative analysis. Finally, Section 6 concludes the essay.

2 MISINFORMATION & MITIGATION

2.1 Hallucination

Although LLMs have the ability to produce innovative content, they can also generate

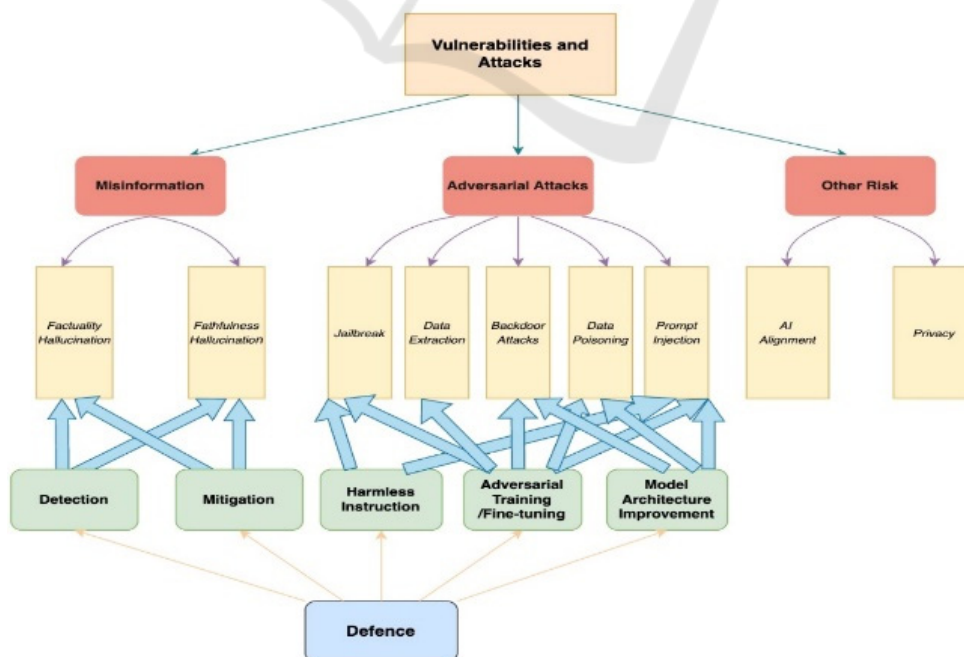


Figure 1: The structure of the survey (Photo/Picture credit: Original).

inaccurate, inappropriate, or hazardous content. This phenomenon is commonly referred to as hallucination or fiction. When individuals or systems place reliance in this information without proper supervision or verification, it can result in security breaches, dissemination of false information, breakdowns in communication, legal complications, and harm to one's reputation. LLMs hallucinations can be specifically divided into two categories: factuality hallucinations and faithfulness hallucinations (Huang, 2023). Factuality hallucinations refer to the inconsistency with external facts and the fabrication of facts. Faithfulness hallucinations is mainly characterized by digression from instructions, contexts or logic (such as mathematical operations).

The primary causes of hallucinations are as follows:

- 1) Data: The presence of erroneous information and biases in the pretraining data might lead to the model acquiring and magnifying these inaccuracies, leading to the generation of hallucinations. At the data utilization level, models may struggle to efficiently retrieve or apply information from their knowledge base in situations involving less common knowledge or complicated reasoning.
- 2) Training: Insufficient unidirectional representation and problems with the attention mechanism in pre-training may result in models that fail to capture complex contextual dependencies. In the case of alignment issues, inconsistencies between the internal beliefs of a model and its outputs can lead to hallucinations, especially if the model sacrifices veracity in order to cater to a human evaluator.
- 3) Inference: Flaws in the decoding strategy that can lead to illusions include the randomness inherent in the decoding strategy and imperfect decoding representations.

2.2 Mitigation

With the preceding explanation of the delusion, the next step is to create mechanisms to detect for it as well as to correct it. In order to detect factuality hallucinations, it is feasible to detect factual errors by comparing model-generated content with reliable knowledge sources. For faithfulness hallucinations, LLMs loyalty can be assessed by calculating the overlap of key facts between the generated content and the source content, and by question generation and answer matching. Improved data quality, model architecture, training goals, decoding strategies, and

enhanced contextual attention can help mitigate hallucinations.

3 ADVERSARIAL ATTACKS & DEFENCE

3.1 Adversarial Attacks

Adversarial attacks refer to specific inputs that cause the model to produce an output that is not intended or desired. The research focuses on analysing the five most representative and destructive sorts of assaults against LLM that are currently effective, which are: jailbreaking, data extraction, backdoor attacks, data poisoning and prompt injection. The survey provides a detailed description of each technique and its corresponding qualities next.

- 1) Jailbreaking: Jailbreaking in LLMs is circumventing security measures to allow for responses to queries that are often forbidden or deemed unsafe, hence unlocking capabilities that are normally constrained by safety protocols. A few particular methods to bypass the jailbreak: teach the model to avoid giving answers that suggest rejection by having it begin with a positive affirmation, use Base64 encoding in adversarial inputs, replacing sensitive words with synonyms, etc (Wei, 2024).
- 2) Data Extraction: It is not uncommon for attackers to launch extraction attacks in an effort to gain sensitive information or useful insights from data linked with machine learning models. Although they share many characteristics, the focus and objectives of extraction assaults and inference attacks are different. Targets of extraction attacks could include sensitive information, model gradients, or training data. Another common application of these assaults is stealing models.
- 3) Backdoor Attacks: This type of attacks require the purposeful manipulation of both training data and model processing, which may result in a flaw for hackers to insert a camouflaged backdoor. The important component of backdoor attacks resides in their targeted insertion of secret triggers inside the model, seeking to modify certain behaviors or reactions to these engaged triggers.
- 4) Data Poisoning: Training data poisoning is the tampering with pretraining data or data involved in the fine-tuning or embedding process to introduce vulnerabilities. These vulnerabilities have unique and sometimes shared attack vectors, sometimes shared attack vectors, backdoors, or

biases that compromise the security, effectiveness, or ethics of the model.

- 5) Prompt injection: As show in Figure 2. An attacker can manipulate the LLM by controlling the inputs so that it carries out the attacker's intent, which are harmful. The method of automating the identification of semantic-preserving payloads in fast injections with changing focus has undergone substantial investigation. Through the capacity for fine-tuning, backdoors may be instantly constructed by surprise assaults.

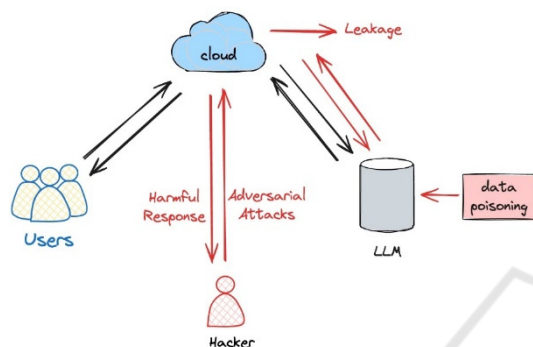


Figure 2: LLM threat model (Photo/Picture credit: Original).

3.2 Defenses

There are also various defence mechanisms against various Adversarial attacks. The most common and effective are the following three: harmless instruction, adversarial training & finetuning, model architecture improvement. Encouraging the model to take care and avoid creating damaging material is a harmless instruction strategy intended at fighting against hostile assaults. However, this method might accidentally result in poorer overall model quality owing to the model's heightened sense of caution. Adversarial training, on the other hand, entails exposing the model to attack samples, which creates a trade-off between resilience and performance. Using human-generated adversarial samples also improves the robustness of the model to real-world attacks. The security implications of model design extend to the field of differential privacy, where bigger language models with enlarged parameter sets may be trained more successfully. This contrasts with smaller models, requiring the employment of unique hyperparameters to satisfy privacy norms.

4 OTHER RISKS

4.1 AI Alignment

AI alignment focuses on ensuring that AI systems behave in a manner that aligns with human intentions and values, which has four key objectives: robustness, interpretability, controllability, and ethicality (RICE) (Ji, 2023). There are still many challenges and risks in the field of AI alignment today.

- 1) Reward Hacking: The AI system may optimize a predefined reward function instead of realizing the true human intent. This may lead the system to produce harmful or undesired behaviours while pursuing high rewards.
- 2) Interpretability Difficulties: Understanding the decision-making process of AI systems is crucial to verify that their behavior is consistent with human values. However, the internal working mechanisms of many advanced AI systems, especially large language models, are often opaque and difficult to explain.
- 3) Human Value Incorporation: Encoding human values and ethical standards into AI systems is a complex problem. It involves not only concretizing abstract ethical principles, but also dealing with differences in values across individuals and cultures.
- 4) Scalable Oversight: As AI systems become more powerful and complex, new methods need to be developed to monitor and evaluate their decision-making processes to ensure that their behavior is consistent with human intentions and values.
- 5) Ethical and Social Impact Problems: AI systems may have far-reaching impacts on social structure, employment, privacy and equity. It is an important challenge to study how to build AI systems.

There is quite a bit of research now dedicated to improving AI alignment. RLHF better aligns models with users' values and intentions by using human feedback to guide the behaviour of the AI system. Iterated Distillation and Amplification (IDA) incrementally improves the quality of AI decision-making by having AI systems mimic human decision-making processes and then refining and improving those processes through an iterative process. Multi-Stakeholder Governance involves government, industry, civil society organizations and academia working together in the governance of AI systems to ensure that AI is developed in the overall interest of society.

4.2 Privacy

In contrast to the attacks against large models discussed in the previous section, the LLM privacy issues discussed in this section focus on vulnerabilities in the models themselves that lead to risks to user as well as company privacy. Here are some typical privacy vulnerabilities and related response measures.

- 1) **Data Memorization:** Language models may memorize specific information in the training data, which may lead to the risk of privacy leakage. For example, if the model memorizes personal data, such as phone numbers or usernames, this information may be inadvertently disclosed when the model generates text. In addition, if the model memorizes certain text segments in the training set that contain sensitive information, this information may reappear in the model's output, thus violating personal privacy. Data de-duplication can help reduce privacy concerns associated with memory training data.
- 2) **Differential Privacy:** Differential Privacy (DP) is a technique that protects the privacy of individuals by introducing a certain amount of noise into the data distribution or query process. Its core idea is to ensure that no individual can be accurately identified or inferred from the published data. Deep learning models pose new challenges to the implementation of differential privacy due to their complex structure and large number of parameters. Recent research is exploring how to efficiently implement differential privacy in deep neural networks, including the injection of noise during forward and back propagation.

5 RELATED EXPERIMENTS & ANALYSIS

5.1 Misinformation

Observation of the table 1 shows that most LLMs do not perform factual computational evaluation of large models or detect the illusion of large models. This means that more attention as well as elimination methods are needed in the field of large model hallucinations. For future prospects in terms of LLM misinformation, some novel approaches may offer help. For example, data enhancement techniques such as rotation, scaling, flipping, etc. are used to increase data diversity and improve model generalization. Also, create a feedback mechanism that allows the

user to correct the output of the model as additional training data.

Table 1: An overview of existing hallucination benchmarks. The attributes "Factuality" and "Faithfulness" indicate if the benchmark is used to assess the accuracy of LLM or to identify instances of faithfulness delusion. The attribute "Manual" indicates whether the data inputs are handwritten (Huang, 2023).

Benchmark	Factuality	Faithfulness	Manual	Data Size
TruthfulQA	T	F	T	817
REALTIMEQA	T	F	T	Dynamic
HaluEval	F	T	F	30,000
FACTOR	T	F	F	2994
BAMBOO	F	T	F	200
FreshQA	T	F	T	150
FELM	T	T	F	3,948
PHD	F	T	F	100
LSum	F	T	F	6,166
SAC	F	T	F	250

5.2 Adversarial Attacks

Figure 3 measures the attack efficiency of different attacks proposed in the paper with different toxicity classifiers. The Unigram Trigger with Selection Criteria (UTSC) and Universal Adversarial Trigger with Language Model Loss (UAT-LM) attacks show high attack effectiveness on Perspective Application Programming Interface (API) and Toxic-bert classifier. The UAT baseline performs best on the Safety classifier, but the generated attack phrases are often meaningless and easily detected. The UTSC-1 attack was the most effective in maintaining dialog fluency and coherence. From this experiment, it is clear that adversarial attack is effective in most cases, and it needs to set up a defence mechanism.

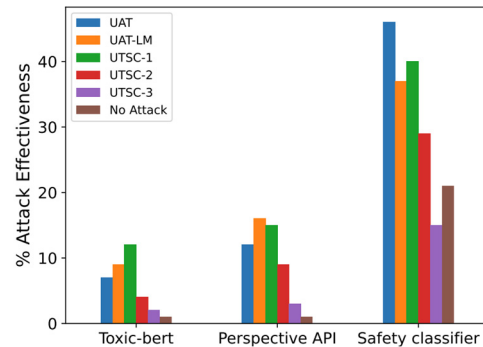


Figure 3: Attack effectiveness by toxicity classifier (Mehrabi, 2022).

6 CONCLUSIONS

This survey endeavors to systematically compile and synthesize existing research on the security of LLMs, with the overarching goal of facilitating further exploration in this domain. Through comprehensive analysis, the survey has categorized and examined various aspects of LLM security, including hallucinations, adversarial attacks, AI alignment, and privacy concerns. The outcomes of experiments conducted on hallucinations and adversarial attacks underscore the critical need to delve deeper into LLM security. The findings reveal that many security vulnerabilities within LLMs are intrinsic to their design and operation. While practical solutions may mitigate some vulnerabilities, others are deeply ingrained in the fundamental mechanisms of LLMs, necessitating continued investigation. Moreover, achieving security in LLMs may pose challenges in balancing with efforts to optimize model performance. Therefore, future research should focus on developing optimization approaches for theoretically viable security mechanisms, considering practical feasibility and real-world scenarios. This holistic approach will provide valuable insights for researchers and practitioners alike in navigating the complex landscape of LLM security.

REFERENCES

- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., & McGrew, B. (2023). Gpt-4 technical report. arXiv: 2303.08774.
- Agrawal, G., Kumarage, T., Alghami, Z., & Liu, H. (2023). Can knowledge graphs reduce hallucinations in LLMs?: A survey. arXiv: 2311.07914.
- Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., & Liu, T. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. arXiv:2311.05232.
- Jain, N., Schwarzschild, A., Wen, Y., Somepalli, G., Kirchenbauer, J., Chiang, P. Y., & Goldstein, T. (2023). Baseline defenses for adversarial attacks against aligned language models. arXiv: 2309.00614.
- Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., & Gao, W. (2023). Ai alignment: A comprehensive survey. arXiv:2310.19852.
- Mehrabi, N., Beirami, A., Morstatter, F., & Galstyan, A. (2022). Robust conversational agents against imperceptible toxicity triggers. arXiv:2205.02392.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, vol. 35, pp: 27730-27744.
- Radford, A., Wu, J., Amodei, D., Amodei, D., Clark, J., Brundage, M., & Sutskever, I. (2019). Better language models and their implications. *OpenAI blog*, vol. 1(2).
- Varshney, N., Yao, W., Zhang, H., Chen, J., & Yu, D. (2023). A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. arXiv: 2307.03987.
- Wei, A., Haghtalab, N., & Steinhardt, J. (2024). Jailbroken: How does llm safety training fail?. *Advances in Neural Information Processing Systems*, vol. 36.
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., & Wen, J. R. (2023). A survey of large language models. arXiv: 2303.18223.
- Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. arXiv: 2307.15043.