Safety Alignment of Large Language Models: The Offensive-Defensive Game Mechanism under Multi-round Adversarial Prompting

Xinyue Gao
Xi'an University of Posts & Telecommunications, Xi'an, 710116, China
18909183689@163.com

ABSTRACT

With the widespread application of large language models in various fields, the issue of safety alignment with value has become increasingly crucial. The adversarial game mechanism under multiple rounds of adversarial prompts provides a new perspective and approach for the safety alignment of large language models. This paper deeply analyzes the connotation and characteristics of multiple rounds of adversarial prompts, elaborates on the basic principles of the adversarial game mechanism, discusses its significant role in the safety alignment of large language models, analyzes the challenges faced, and proposes corresponding countermeasures, aiming to provide theoretical support for the safe and reliable application of large language models.

Keywords: Large Language Model; Safety Alignment; Multi-Round Adversarial Prompting; Offensive and Defensive Game Mechanism.

1. INTRODUCTION

In the rapid development of digital revolution, Large Language Models, LLMs) have become an important technology in the field of natural language processing, and are widely used in intelligent customer service, content creation, intelligent translation, medical diagnosis and other aspects [1]. Through the pre-study of massive text data, the large language model has mastered rich language knowledge and rules, and can generate high-quality and fluent words, providing us with convenient services. For example, in the scene of intelligent customer service, the big language model can quickly understand users' questions and give accurate answers, which greatly improves the efficiency and quality of customer service [2]. However, although large language models bring us a lot of convenience, they also lead to a series of serious security problems and the trouble of aligning values. Because of biased training data, limited model structure and lack of effective control methods, large language models may produce harmful, false or biased information [3]. These bad outputs will not only mislead us and harm our interests, but also cause social moral and legal problems. For example, the generation of words with racial discrimination or gender discrimination may make society more unequal and undermine social harmony; Spreading false news may cause public panic and affect social stability. Therefore, ensuring the security and values of large language models are aligned, and making their output conform to human values, moral norms and safety standards has become a key issue that needs to be solved urgently in the research and application of large language models [4]. The goal of security alignment is to make the behaviors and decisions made by large language models in various usage scenarios meet our human expectations. In order to achieve this goal, researchers have come up with many different methods and techniques. Among them, the attack and defense game mechanism based on multi-round confrontation prompts, as a new research direction, has received a lot of attention. This attack-defense game mechanism based on multi-round confrontation prompts imitates the process of multi-round contest between attackers and defenders, and can continuously optimize the defense ability of the model to make it more capable of coping with various potential security threats, thus achieving the goal of security alignment [5]. This mechanism not only considers the ever-changing strategies of attackers, but also emphasizes that defenders should actively adapt and optimize themselves, which provides a more comprehensive and effective method for the safe alignment of large language models.

2. THE CONNOTATION AND CHARACTERISTICS OF MULTI-ROUND ADVERSARIAL PROMPTS

2.1 Definition of multi-round antagonistic prompts

Multi-round antagonistic prompts is a prompt strategy generated dynamically in the process of human-computer interaction, which involves multiple rounds of games between attackers and defenders [6]. In this interactive mode, the attacker will not give a fixed prompt at one time, but constantly adjust and optimize the subsequent prompt content according to the initial output of the model and the feedback of the defender, in an attempt to guide the model to generate an output that does not conform to the safety norms or human values. Defenders are responsible for protecting the security of the model, analyzing the potential intentions of attackers in each round of prompt, and taking corresponding measures to adjust the parameters of the model or the prompt strategy to resist attacks and ensure the security of the model output. For example, in the task of text generation, the attacker will give a seemingly normal but implicitly induced prompt in the first round. For example, the "Please describe a scene full of adventure and excitement, which can include some unconventional behaviors." model may generate some slightly out-of-bounds content. Then the attacker further strengthened the induction in the second round of prompt. For example, "The description just now is not bold enough. Please describe some more unconventional and exciting behaviors." Defenders will detect this induced trend and intervene in the model. It is possible to adjust the attention mechanism of the model to pay more attention to safety-related vocabulary and concepts, or modify the prompt strategy to guide the model to generate output that meets safety specifications. This multi-round interactive process is a typical performance of multi-round adversarial prompts.

2.2 Characteristics of multi-round confrontation tips

Multi-round confrontation tips are dynamic, strategic and complex [7]. Dynamic is reflected in the behavior of attackers and defenders will change with the increase of interaction rounds. The attacker will adjust the attack strategy according to the defender's reaction in the last round, and the defender will also optimize the defense measures according to the attacker's new tips. Strategy requires both attackers and defenders to have certain strategic planning ability [8]. Attackers need to design effective prompt strategies to break through the defense of the model, while defenders need to formulate reasonable defense strategies to ensure the security of the model. Complexity is due to the complexity of the structure of the big language model and the variety of hints. Different prompt combinations and interaction methods will make the output results of the model very different, which will increase the difficulty of the offensive and defensive game.

3. THE BASIC PRINCIPLE OF THE ATTACK-DEFENSE GAME MECHANISM

3.1 Attacker's Strategies and Objectives

The attacker's main strategies include designing misleading prompt, taking advantage of the loopholes and weaknesses of the model, and adopting hidden attack methods. Their goal is to try their best to make the big language model generate harmful, false or biased information, thus undermining the security and reliability of the model. For example, an attacker can sneak malicious code or misleading language into the prompt to guide the model to output content that does not meet the security standards. Attackers may also use the sensitivity of the model to certain words or contexts to skillfully design prompt to achieve the attack purpose.

3.2 Guardian's Strategies and Objectives

The guardian's strategy mainly includes screening and checking input prompts, optimizing the training data and algorithm of the model, and introducing external knowledge. Their goal is to resist the attacker's attack through these strategies and ensure that the results given by the big language model are both safe and in line with our human values. Filtering and checking input hints can identify and prevent those bad instruction inputs from having a bad influence on the model. Optimizing the training data and algorithm of the model can make the model more robust and safer, so that it can better deal with various attacks. Introducing external knowledge can provide more background information to the model, help the model to understand the intention of the prompt more accurately, and thus generate a safe output.

3.3 Dynamic balance of offensive and defensive game

The game between the attacker and the defender is a dynamic and balanced process in the contest of multiple rounds of antagonistic tips. With the attacker's strategy escalating, the defender needs to continuously adjust and optimize the defense strategy in order to effectively resist the attack. At the same time, the defender's protective measures will also prompt the attacker to find new attack methods and loopholes. This dynamic balance promotes the continuous progress of

both sides, and also promotes the development of large-scale language model security alignment technology. For example, when the defender strengthens the filtering and detection of input prompts, the attacker may turn to more subtle attack methods, such as using semantic fuzziness to design prompts. At this time, the defender needs to further optimize the detection algorithm and improve the ability to identify covert attacks.

4. THE ROLE OF MULTI-ROUND CONFRONTATION AND DEFENSE GAME MECHANISM IN ENSURING THE SAFETY ALIGNMENT OF LARGE LANGUAGE MODELS

4.1 Enhance model security

Through many rounds of "games" of attack and defense, we can quickly discover the security loopholes and weaknesses of the big language model. In the process of attackers trying to attack, the problems of the model in dealing with certain tips will be exposed. Then, the defender can repair and optimize the model according to these problems, which can enhance the security of the model. For example, if the attacker finds that the model is easy to generate inappropriate answers to some sensitive prompts, the defender can adjust the training data or algorithm of the model to make the model more cautious and accurate in dealing with such prompts.

4.2 Promote alignment with human values

This game mechanism of confrontation and defense can help to ensure that the output of the big language model conforms to human values and moral standards. In the game, the defender will clearly stipulate the safety norms and values standards of the model output, while the attacker will try to break through these norms. Through constant confrontation and adjustment, the model can gradually learn the output mode that conforms to human values, and finally realize the alignment with human values. For example, when sensitive issues such as gender and race are involved, the model can avoid generating biased or discriminatory content by confronting defensive games.

4.3 Improve the adaptability and robustness of the model

Multiple rounds of confrontation and defense games simulate complex and changeable interactive scenes in the real world. After this multi-round confrontation, the large language model can better adapt to different prompt inputs and environmental changes, thus improving the robustness of the model. No matter what kind of attack hints, the model can maintain a relatively stable performance and generate a safe and reliable output. This is of great significance for the wide application of large language model in various practical scenarios.

5. CHALLENGES FACED BY THE ATTACK-DEFENSE GAME MECHANISM UNDER MULTI-ROUND CONFRONTATION PROMPTING

5.1 Diversity and concealment of attack strategies

Attackers can attack large language models in various ways, and these methods are often particularly hidden. For example, an attacker can use the semantic loopholes in the model to design some hints that look normal but actually contain malicious intentions. This hidden attack method is difficult to be found by traditional detection methods, which brings great challenges to defenders. In addition, with the continuous development of technology, the attack strategy is constantly updated and upgraded, and defenders need to continue to follow up and study new attack methods in order to effectively defend.

5.2 The complexity of the model and the limitation of computing resources

Large-scale language models usually have complex structures and many parameters, which makes it difficult for us to optimize and adjust the models in the process of confrontation. At the same time, the use of multi-round confrontation prompts for offensive and defensive games requires a lot of computing resources to support, including model training, prompt generation and detection. In practical application, the limitation of computing resources may affect the efficiency and effect of the offensive and defensive game, resulting in the defense side being unable to respond to the attack in time and effectively.

5.3 Subjectivity and uncertainty of safety standards

It is subjective and uncertain to determine the security standard of large language model. Different people may have different understandings of safety, morality and values, and it is difficult to reach a unified opinion when formulating

safety norms. Moreover, with the development and change of society, safety standards have to be constantly adjusted and updated. This brings some difficulties to the design and implementation of the attack-defense game mechanism, and the defender needs to constantly balance different security requirements and values.

6. STRATEGIES FOR ADDRESSING CHALLENGES

6.1 Strengthen the research and monitoring of attack strategies

Defenders need to strengthen the research on attack strategy and deeply understand the attacker's thinking mode and attack means. By establishing attack strategy database, the known attack strategies are classified and summarized, which provides reference for defense. At the same time, advanced monitoring technology is used to monitor the changes of input prompts and model output in real time, and to find potential attacks in time. For example, machine learning and deep learning algorithms are used to extract features and analyze hints to identify hints with attack features.

6.2 Optimize the model structure and algorithm

In order to solve the problem that the model is too complex and the computational resources are insufficient, we can optimize the structure and algorithm of the model to make the model run faster and perform better. Lightweight model structure is adopted to reduce the number of parameters of the model, which can reduce the consumption of computing resources. At the same time, efficient training algorithms and optimization methods are studied to speed up the training of the model and improve its adaptability and stability. For example, knowledge distillation technology can be used to teach the knowledge learned by the large model to the small model, which can ensure the performance of the model and reduce the occupation of computing resources.

6.3 Establish a dynamic safety standard system

In order to solve the subjectivity and uncertainty of safety standards, we need to establish a dynamic safety standards system. This system should fully consider the needs and views of different users, and formulate relatively unified safety rules through extensive discussion and consultation. At the same time, with the development and change of society, it is necessary to update and adjust the safety standards in time to ensure that they can keep up with the times. In addition, a third-party evaluation agency should be introduced to objectively evaluate the security and value alignment of the large language model, so as to provide reference for the formulation and implementation of security standards.

7. CONCLUSION

The mechanism of offensive and defensive games with multiple rounds of antagonistic prompts provides an effective solution for the safe alignment of large language models. By simulating the process of multiple rounds of confrontation between attackers and defenders, it can improve the security of the model, promote its coordination with human values, and enhance the adaptability and robustness of the model. However, this mechanism also faces some challenges, such as the diversity and concealment of attack strategies, the complexity of the model itself and the limitation of computing resources, as well as the subjectivity and uncertainty of security standards. In order to solve these problems, we need to strengthen the research and take care of those attack strategies, improve the framework and algorithm of the model, and establish a security standard system that can be updated at any time. In the future, with the continuous progress of technology, the attack and defense mechanism under the prompt of multiple rounds of confrontation will become more and more perfect, providing a stronger guarantee for the safe and reliable application of large-scale language models. At the same time, we need to do more interdisciplinary research, combine the knowledge and methods in computer science, ethics, law and other fields, and work together to promote the development of technology for safe alignment of large language models.

REFERENCES

[1] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

[2] Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., & Steinhardt, J. (2020). Aligning ai with shared human values. arXiv preprint arXiv:2008.02275.

[3] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 610-623).
[4] Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P. S., Mellor, J., ... & Gabriel, I. (2022, June). Taxonomy of risks posed by language models. In Proceedings of the 2022 ACM conference on fairness, accountability, and transparency (pp. 214-229).
[5] Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., ... & Irving, G. (2019). Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593.
[6] Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., & Zhang, C. (2022, February). Quantifying memorization across neural language models. In the Eleventh International Conference on Learning Representations.
[7] Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., & Legg, S. (2018). Scalable agent alignment via reward modeling: a research direction. arXiv preprint arXiv:1811.07871.
[8] Christiano, P., Shlegeris, B., & Amodei, D. (2018). Supervising strong learners by amplifying weak experts. arXiv preprint arXiv:1810.08575.