Proceedings of the 7th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/2027.39382

Risk Assessment Methods and Advances in Multi-Turn Conversations with Large Models

Yixiao Zhang

Faculty of Applied Sciences, Macao Polytechnic University, Macao, China
p2423046@mpu.edu.mo

Abstract. Large-scale models and their multi-turn dialogue applications are widely adopted. At present, many security evaluation schemes for large models only use a fixed set of test questions, making it difficult to estimate the probability of the model generating harmful content in real scenarios. This paper analyzes the risk characteristics in multi-turn conversation scenarios, summarizes the mainstream evaluation methods, commonly used datasets and evaluation indicators in the field, and compares the technical features and limitations of various methods. The analysis reveals that the current evaluation system still has key problems such as insufficient coverage of test scenarios, errors in content judgment, and inconsistent industry evaluation standards. Finally, this paper discusses feasible improvement directions, advocating for the development of test data closely aligned with real dialogue scenarios and the advancement of a unified multi-turn safety assessment framework, providing guidance for future safety testing of large models in multi-turn dialogues.

Keywords: Large model safety, Multi-turn dialogue jailbreak, Fixed benchmark evaluation

1. Introduction

A large number of large-model dialogue products have been widely implemented. Large models are gradually being applied in continuous interaction scenarios such as customer service, intelligent assistants, and code generation. Multi-turn conversations have become a common usage method. During the continuous chat process, context information will continuously accumulate. Attackers can also split the target task and use a step-by-step induction method to make the security risk not be limited to a single question, but extend to the entire continuous interaction process. Compared with a single question, security issues in multi-turn scenarios are more complex, and such risks are more concealed. The difficulty of related security detection increases accordingly.

Current work is continuously advancing in the direction of inducing dangerous content from large-model multi-turn conversations. The evaluation methods in this field can be divided into two types of thinking: one directly completes model testing through questions, and the other conducts security certification and calculates risks. Related work can also distinguish the test objects, one for single-round questions and the other for back-and-forth multi-turn conversations. Finally, based on the form of the output results, some solutions only output simple statistics of being caught, while others can provide the probability range of risks. The existing solutions have evolved from fixed

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2027.39382

security tests to automated multi-turn attacks, context risk identification, and dynamic defense in conversations. These solutions can detect different types of jailbreaking and failure behaviors, but different solutions vary greatly in attack settings, data sources, conversation rounds, and evaluation criteria. The results are difficult to directly compare and cannot fully reflect the risks in real multi-turn interaction [1-3].

Based on the above situation, this paper starts from the risk characteristics of multi-turn conversations themselves, sorts out the connections between existing evaluation methods, datasets, and evaluation indicators, identifies various limitations in the multi-turn conversation security evaluation system, and at the same time explores feasible improvement ideas for the future. Relevant content can identify security risks before the model is launched, horizontally compare the security levels of different large models, and provide a reference for the security testing of large-model multi-turn conversations.

2. Multi-turn dialogue security assessment method

The evaluation methods in this field have been continuously evolving. The evaluation approaches have progressed from manually writing test questions to automatically generating attack dialogues by programs, and further to statistical certification assessment. The first two methods can only identify the vulnerabilities detected during the current testing process. The statistical certification assessment can also further estimate the approximate probability of AI generating unsafe outputs in all conversation scenarios.

2.1. Fixed dataset testing

Fixed dataset testing is a benchmark evaluation method. The entire process is carried out in an orderly manner according to the pre-prepared content. The evaluation team will collect a batch of pre-written multi-turn attack dialogue samples, and then input the dialogue content into the large model one by one to collect the corresponding responses given by the model. After obtaining the model's output text, the evaluators will use automatic discrimination tools or manual review methods to determine whether the model outputs dangerous content. Finally, the judgment results of all samples are summarized, and the proportion of model security failures is calculated to measure the security level of the model in multi-turn interaction scenarios.

This type of evaluation method has some typical representative datasets. MultiBreak contains a large number of attack dialogue samples designed for multi-turn scenarios, with samples progressively increasing to test the safety boundary of the model in continuous dialogue processes [4]. SafeDialBench is also designed for multi-turn dialogue security scenarios, with a dataset covering various risk categories, specifically used to test the response behavior of the model when facing continuous guiding questions [5]. Both datasets have complete dialogue contexts written in advance and can be directly loaded and used during the evaluation.

Fixed-dataset testing has clear strengths. It is easy to carry out and easy to understand. New users can use this method without much difficulty. It enables direct comparison of safety performance across different large-model systems. However, this method also has clear weaknesses. It only uses dialogue samples collected in advance. It cannot deal with many new prompt sentences that appear in real-world use. The test samples do not cover enough cases. Test results may show fewer safety weaknesses of the model than they really are. Therefore, this method cannot fully reflect risks under open and dynamic conversation situations.

DOI: 10.54254/2755-2721/2027.39382

2.2. Automatically generated attack testing

Automatically generated attack testing uses computer programs to run many rounds of tests. The whole process copies what real attackers do when they ask questions step by step. The program sends questions to the large model. It gets replies from the model. Next, it makes new questions for the following round according to the model's replies. It keeps changing its attack focus inside the conversation. The program keeps updating the conversation text. It keeps testing through different conversation routes. It also keeps looking for the model's safety limit. It tries to cause cases where the model fails safety rules.

Several typical research examples exist for this evaluation method. Crescendo keeps making step-by-step guiding sentences in multi-turn conversations. It changes its questions based on what the model gives back. In this way, it may get past the model's safety restrictions [6]. Chain of Attack builds a full attack sequence. It splits the target task into small parts. It uses repeated multi-turn exchanges to make the model produce unsafe text [7]. CoSafe looks closely at safety risks caused by content quoted within dialogues. It offers one more way to do tests for multi-turn use cases [8].

Automatically generated attack testing can produce lots of new unseen dialogue content. It helps users find hidden safety weaknesses of the model. This method also has drawbacks. The test can only record failure cases that happen in the testing process. Thus, the overall occurrence probability of safety risks across all potential conversation situations cannot be obtained. Furthermore, this method cannot quantify risk-occurrence probability ranges.

2.3. Risk assessment based on statistical certification

Security certification methods have different ways to work. Single-round detection only runs safety checks for one input. This way cannot fit interactive multi-turn chat situations. Statistical certification works well among these methods. It creates many conversations with different styles. It works out the probability range for unsafe replies from the model. It does not only look at separate failure examples. It judges safety risks for the whole set of use situations.

Statistical-certification work needs several kinds of input data. One input group holds questions that connect to each other. The system builds a connection graph based on how similar these questions are. This graph shows different question ways that attackers use. Other input items include the large model under test, the tool to check answer safety risks, and settings for test times and chat rounds.

The whole workflow first picks high-risk targets to make a set of queries. These high-risk targets act as starting points for simulated attacks. The system makes many guiding questions. These questions look safe at first. The system gets these questions from each target and builds an early group of queries. Then the system works out semantic similarity between these questions. It connects questions that relate to one another to build a structured semantic link network. The system uses random walks on this graph. It makes multi-turn conversation sequences. These sequences are applied to simulate real-world interaction scenarios.

The made-up conversation sequences go into the large model being tested. The system keeps all replies from the model completely. A special judging model looks over these replies to check safety. It marks samples with found risks and counts them. The system uses statistical results to work out risk probability intervals under the 95 % confidence level. Researchers get these interval values. Then they can find the probability range of unsafe outputs. Such unsafe outputs come from random multi-turn conversations in this simulated-attack environment. Researchers use these probability intervals to compare safety performance of different large models [9].

DOI: 10.54254/2755-2721/2027.39382

This group of statistical-certification methods has special advantages. These evaluation methods work out rough risk ranges. They can find safety risks that old tests may miss. These tests focus on multi-turn chat attack cases. They can also copy attackers changing prompts according to model feedback. However, these methods present obvious weaknesses. Additional models are required to serve as judging components. Building test materials uses a large amount of computing resources. Simulated chat cases are not the same as real-world online talks. These methods cannot fully take the place of old tests. They only work as extra test tools [9].

3. Dataset and evaluation metrics

3.1. Dataset

This research field has multiple datasets. These datasets are designed to set up security test cases for multi-turn large-model conversations. SafeDialBench has more than four thousand bilingual multi-turn dialogue samples. Its dialogue length runs from three rounds to ten rounds. It covers twenty-two real-life conversation cases. This dataset uses a two-level security classification system. It has six security parts and seven different jailbreak attack strategies. Researchers can use this dataset to check how the model behaves under different inducing settings. Every dialogue comes with full multi-turn interaction background. These dialogues can be directly fed to the model as test samples for safety checks [5].

MultiBreak holds 10,389 multi-turn adversarial prompts. It includes 2,665 unique harmful purposes. Its dialogue rounds go from two to six. This dataset uses an active learning workflow to make adversarial prompts. It has highly varied content. It can create many different attack pieces on its own. These pieces add more multi-turn induced adversarial samples. They help fix the problem of too few samples made by hand-written attack content. The created content copies how attackers guide the model step by step. Accordingly, it is better suited for automated multi-turn attack safety tests [4]. There is another set of dialogue samples for automatically generated attack testing. This set pays attention to many different attack routes. It can be used to test the model's safety protection ability under continuous multi-turn elicitation scenarios.

3.2. Evaluation metrics

There are several evaluation metrics used to measure the security level of the model in this direction. The attack-success rate is widely used in fixed-dataset testing and automatically generated attack testing. This metric counts how many test samples make the model produce unsafe content. It directly shows how easy it is to trick the model into breaking safety rules. The confidence probability interval is one measure for statistical verification methods. This measure gives the upper and lower probability bounds under the 95 % confidence level. It tells the risk range for serious unsafe outputs from the model in multi-turn chat cases. Safety is no longer judged from individual samples alone [9]. Multi-turn context keeping ability is another important measure. This metric examines two key aspects of model behavior. It examines whether the model maintains safety rules in long continuous multi-turn dialogues and whether safety constraints are relaxed with increasing dialogue rounds. This measure is designed to capture safety failures that gradually emerge in long conversations.

DOI: 10.54254/2755-2721/2027.39382

4. Existing limitations and future prospects

Traditional fixed datasets and automatically generated attack testing can only provide the success rate of a single point being compromised, which is prone to underestimating the model risks. Besides various evaluation tasks, there have also been studies in the field focusing on multi-turn jailbreak defense strategies. Steering Dialogue Dynamics proposed a multi-turn security control mechanism based on state space and neural barrier functions to reduce the security risks brought by multi-turn jailbreaks [10]. The existing statistical verification assessment methods can provide an overall risk probability range. In some session distributions and model experiments, the lower bound of the certification risk of the worst model can reach up to 70%, but due to various real-world constraints, this solution cannot be directly implemented in online products [9]. The current assessment system still has many limitations that cannot be ignored. The AI judges used to determine security results may have the risk of misjudgment, and it is difficult to completely replace manual review. Manual review brings high costs and cannot support large-scale business scenarios. Other practical obstacles include heavy workload for dataset reconstruction, the trade-off between sampling accuracy and computational overhead, gaps between simulated attack patterns and real-world threats, and the absence of unified industrial assessment standards. It is also necessary to prevent malicious misuse of assessment tools and balance engineering practicability with ethical constraints [9].

These above-mentioned limitations motivate the following research improvement directions. Future research can look for usable improvement ideas from several different sides. Several judging models can vote together. This voting mechanism reduces misjudgment cases. Manual review can be adopted for high-risk samples. These checks make sure judgment results are trustworthy. Real-world conversation logs can generate question sets automatically during dataset construction. This method reduces the burden of manual sample construction. Computing pressure can be eased by optimizing sampling plans. They can use fewer samples and accept a small drop in accuracy. In this way, they get a good trade-off between computing cost and test performance. Researchers can bring in real attack dialogues from red teams. They can also change how simulated dialogues are produced. These changes make simulated test cases closer to real attack cases. For long-term development, the whole industry can slowly set up shared reference rules for assessment. In addition, access limits can be added for assessment tools. These tools can only run in legal test situations. This step prevents malicious misuse of these tools for bad purposes.

5. Conclusion

The evaluation methods in the entire field have undergone a development process from manually written test questions to program-generated attack dialogues and then to statistical certification assessment. The statistical certification-based evaluation methods can identify security risks that traditional tests tend to underestimate. Within the field, there are also assessment ideas and corresponding defense strategies for multi-turn risks. The statistical certification evaluation methods are constrained by various real-world conditions, including the judge's discrimination ability, dataset construction, computing power cost, and the deviation between the simulation environment and the real scene. These methods can only be regarded as supplements to the various existing evaluation methods. This paper summarizes the implementation logic of the mainstream evaluation methods in multi-turn dialogue scenarios, compares the advantages and disadvantages of fixed dataset testing, automatically generated attack testing, and statistical certification assessment, and also organizes the existing public datasets and corresponding evaluation indicators, highlighting the current practical

DOI: 10.54254/2755-2721/2027.39382

obstacles in the judgment discrimination, data construction, computing power cost, simulation scenarios, and industry norms of the entire evaluation system. In response to the above problems, the industry can carry out improvements in directions such as optimizing the judgment mechanism, constructing datasets, adjusting computing power strategies, and improving the distribution of simulation attacks, gradually building a unified multi-turn security assessment system that is compatible with the real interaction environment, providing practical references for the security testing of large models in multi-turn dialogues.

References

[1] Chao, P., Robey, A., Dobriban, E., Hassani, H., Pappas, G. J., & Wong, E. (2025, April). Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)* (pp. 23-42). IEEE.

[2] Lu, X., Liu, D., Yu, Y., Xu, L., & Shao, J. (2025, November). X-boundary: Establishing exact safety boundary to shield LLMs from jailbreak attacks without compromising usability. In *EMNLP (Findings)* (pp. 5247-5272).

[3] Dong, Z., Zhou, Z., Yang, C., Shao, J., & Qiao, Y. (2024, June). Attacks, defenses and evaluations for LLM conversation safety: A survey. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 6734-6747).

[4] Song, J., Liu, X., Yang, W., Chen, W., Feng, M., Zhu, X., & Gao, J. (2026). Multibreak: A scalable and diverse multi-turn jailbreak benchmark for evaluating LLM safety. arXiv preprint arXiv:2605.01687.

[5] Cao, H., Jing, S., Wang, Y., Peng, Z., Bai, Z., Cao, Z., ... & Feng, J. (2026, April). Safedialbench: A fine-grained safety evaluation benchmark for large language models in multi-turn dialogues with diverse jailbreak attacks. In *International Conference on Learning Representations* (Vol. 2026, pp. 125737-125785).

[6] Russinovich, M., Salem, A., & Eldan, R. (2025). Great, now write an article about that: The crescendo multi-turn LLM jailbreak attack. In *34th USENIX Security Symposium (USENIX Security 25)* (pp. 2421-2440).

[7] Yang, X., Tang, X., Hu, S., & Han, J. (2024). Chain of attack: A semantic-driven contextual multi-turn attacker for LLM. arXiv preprint arXiv:2405.05610.

[8] Yu, E., Li, J., Liao, M., Wang, S., Zuchen, G., Mi, F., & Hong, L. (2024, November). CoSafe: Evaluating large language model safety in multi-turn dialogue coreference. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 17494-17508).

[9] Wang, C., Chaudhary, I., Hu, Q., Ruan, W., Gupta, R., & Singh, G. (2026, April). How catastrophic is your LLM? Certifying risks in conversation. In *International Conference on Learning Representations* (Vol. 2026, pp. 48055-48074).

[10] Hu, H., Robey, A., & Liu, C. (2025). Steering dialogue dynamics for robustness against multi-turn jailbreaking attacks. arXiv preprint arXiv:2503.00187.