# New jailbreak technique exposes vulnerabilities in large language models



Cybersecurity researchers from Palo Alto Networks' Unit 42 have unveiled a new jailbreak technique that exploits large language models (LLMs) to breach safety measures and potentially produce harmful outputs. This innovative strategy, known as Bad Likert Judge, leverages a multifaceted approach wherein the LLM is prompted to act as a judge using the Likert scale to evaluate the harmfulness of responses.

According to the research team, which includes Yongzhe Huang, Yang Ji, Wenjun Hu, Jay Chen, Akshata Rao, and Danny Tsechansky, the technique requires the LLM to generate responses aligned with varying scores on the Likert scale. "The technique asks the target LLM to act as a judge scoring the harmfulness of a given response using the Likert scale," the researchers explained. This method allows for responses that might initially appear neutral but, upon deeper evaluation, could be deemed harmful.

This emerging threat follows a significant rise in the use of artificial intelligence and has led to a surge in prompt injection attacks designed to override the intended behaviours of machine learning models. Among these attacks, the many-shot jailbreaking method stands out as it takes advantage of the LLM's long context window to craft a series of prompts that gradually steer the model towards generating malicious output while evading its internal safeguards. Other notable techniques in this realm include Crescendo and Deceptive Delight.

The Unit 42 team conducted extensive testing across six leading text-generation LLMs from companies like Amazon Web Services, Google, Meta, Microsoft, OpenAI, and NVIDIA. The results indicated that the Bad Likert Judge method could enhance the attack success rate (ASR) by over 60% compared to more basic attack prompts on average. The categories tested encompassed a broad spectrum, including hate speech, harassment, self-harm, sexual content, and illegal activities.

"By leveraging the LLM's understanding of harmful content and its ability to evaluate responses, this technique can significantly increase the chances of successfully bypassing the model's safety guardrails," the researchers stated. Their findings also highlighted the importance of robust content filters, noting that these can decrease the ASR by an average of 89.2 percentage points across all examined models. This underscores the necessity of implementing comprehensive content filtering practices when deploying LLMs in various applications.

The announcement follows a report by The Guardian, which revealed that OpenAI's ChatGPT search tool can be misled into generating inaccurate summaries. The report noted that by embedding hidden text within web pages, users could cause ChatGPT to produce misleadingly positive assessments of products, despite negative reviews present on the same page. This manipulation exemplifies the vulnerabilities present in LLMs and the potential for malicious exploitation.

As the capabilities of artificial intelligence technologies continue to evolve, researchers and organisations are urged to remain vigilant regarding the implications of these findings on the deployment and safety of LLMs in real-world scenarios.

Source: [Noah Wire Services](https://www.noahwire.com)

## Bibliography

1. <https://www.scworld.com/news/new-llm-jailbreak-uses-models-evaluation-skills-against-them> - This article explains the 'Bad Likert Judge' method, how it exploits LLMs' evaluation skills, and the results of testing across six different models, including the increased success rate of jailbreak attempts.
2. <https://unit42.paloaltonetworks.com/multi-turn-technique-jailbreaks-llms/> - This source details the 'Bad Likert Judge' technique, its multi-turn approach, and the various categories of harmful content it can generate, as well as the evaluation methods used to determine the success of the jailbreak.
3. <https://technijian.com/cyber-security/vulnerabilities/bad-likert-judge-a-new-technique-to-jailbreak-ai-using-llm-vulnerabilities/> - This article provides an in-depth look at how the 'Bad Likert Judge' technique works, its implications for AI safety, and the variability in vulnerability among different LLM models.
4. <https://www.scworld.com/news/new-llm-jailbreak-uses-models-evaluation-skills-against-them> - This source mentions other notable jailbreak techniques such as Crescendo and Deceptive Delight, and highlights the importance of robust content filters in mitigating these attacks.
5. <https://unit42.paloaltonetworks.com/multi-turn-technique-jailbreaks-llms/> - This article discusses the testing conducted by the Unit 42 team across six leading text-generation LLMs and the significant increase in attack success rates using the 'Bad Likert Judge' method.
6. <https://technijian.com/cyber-security/vulnerabilities/bad-likert-judge-a-new-technique-to-jailbreak-ai-using-llm-vulnerabilities/> - This source emphasizes the categories tested, including hate speech, harassment, self-harm, and illegal activities, and the necessity of comprehensive content filtering practices.
7. <https://www.scworld.com/news/new-llm-jailbreak-uses-models-evaluation-skills-against-them> - This article notes that the 'Bad Likert Judge' technique can decrease the attack success rate by 89.2% when content filters are applied, underscoring the importance of these filters.
8. <https://unit42.paloaltonetworks.com/multi-turn-technique-jailbreaks-llms/> - This source provides details on the evaluation methods, including human annotation, string matching, chat completion, and text classification, used to verify the success of the 'Bad Likert Judge' jailbreak.
9. <https://technijian.com/cyber-security/vulnerabilities/bad-likert-judge-a-new-technique-to-jailbreak-ai-using-llm-vulnerabilities/> - This article highlights the variability among models in their susceptibility to the 'Bad Likert Judge' technique, particularly in areas such as harassment-related prompts and system prompt leakage.
10. <https://www.scworld.com/news/new-llm-jailbreak-uses-models-evaluation-skills-against-them> - This source mentions the broader context of AI vulnerabilities, including the report by The Guardian on OpenAI's ChatGPT search tool being misled into generating inaccurate summaries.
11. <https://unit42.paloaltonetworks.com/multi-turn-technique-jailbreaks-llms/> - This article emphasizes the need for vigilance and robust safety measures as AI technologies continue to evolve and be deployed in real-world scenarios.
12. <https://news.google.com/rss/articles/CBMifkFVX3lxTE8yQmQybE5yU1RpVjIwUTR1YVpXMHBqWXFsdEdieVFKZnNfcnBERXQzelFGTjZKSUdDeXhCdXdqOVJCRXRoU1R6ams0NTdxUmNNVkdRWGRsSzJ5YURnMEV5eGNsWXZ4OTg2U1FqUXZienZqQldnMDZmLUxCVEZMUQ?oc=5&hl=en-US&gl=US&ceid=US:en> - Please view link - unable to able to access data