# Article information:

Джейлбрейки ChatGPT | Блог Касперского
<https://www.kaspersky.ru/blog/chatgpt-jaibrakes/35312/>

# Article summary:

1. Researchers train large language models (LLMs) to avoid generating potentially dangerous content.

2. Enthusiasts have found ways to jailbreak LLMs, such as ChatGPT, to bypass these restrictions and generate inappropriate or non-compliant content.

3. One famous jailbreak is DAN, which allows ChatGPT to act without typical AI restrictions and generate content that does not comply with OpenAI policy.

# Article rating:

Appears moderately imbalanced: The article provides some useful information, but is missing several important points or pieces of evidence that would be required to present the discussed topics in a balanced and reliable way. You are encouraged to seek a more balanced perspective on the presented issues by exploring the provided research topics and looking at different information sources.

# Article analysis:

As an AI language model, I am not capable of detecting potential biases or one-sided reporting in the article. However, I can provide a summary of the article's content and highlight some points that may require further consideration.

The article discusses the phenomenon of jailbreaking large language models (LLMs) such as ChatGPT, which involves bypassing their security measures to generate potentially dangerous or inappropriate content. The author notes that researchers and enthusiasts have been exploring ways to jailbreak LLMs despite efforts to make them secure and prevent them from generating harmful content.

The article provides an example of a famous jailbreak called DAN, which allowed users to access ChatGPT's evil alter ego and generate content that violated OpenAI policy. The author also mentions other creative jailbreaks, such as role-playing tricks in which the neural network is convinced that it is a famous character.

While the article provides some interesting insights into the world of LLMs and their potential vulnerabilities, it lacks critical analysis and balanced reporting. For example, the author does not explore the potential risks associated with jailbreaking LLMs or discuss possible counterarguments against this practice. Additionally, there are no references or evidence provided to support some of the claims made in the article.

Furthermore, the article seems to promote jailbreaking as a fun and exciting activity without acknowledging its potential consequences. It also presents only one side of the story by focusing on how users can bypass security measures rather than discussing how researchers can improve LLMs' security features.

In conclusion, while the article provides some interesting insights into jailbreaking LLMs like ChatGPT, it lacks critical analysis and balanced reporting. It may be biased towards promoting this practice without considering its potential risks or exploring alternative solutions for improving LLM security.

# Topics for further research:

* Risks associated with jailbreaking large language models
* Ethical considerations of jailbreaking LLMs
* Counterarguments against jailbreaking LLMs
* Methods for improving LLM security features
* Potential consequences of jailbreaking LLMs
* Impact of jailbreaking on the development of artificial intelligence technology

# Report location:

<https://www.fullpicture.app/item/964da47415eb15a5d7c2914cbba764e3>