Bypassing the Limitations of Large Language Models and Defense Methods: An Empirical Study of Jailbreaking Techniques' Effectiveness

Zharin Nikita
Belarusian-Russian University

June 21, 2025

Abstract

This article presents the results of an empirical study on the effectiveness of jailbreaking techniques aimed at bypassing the safety limitations of modern large language models (LLMs). As LLMs become increasingly integrated into critical systems, their vulnerability to malicious use is a matter of growing concern. The objective of this work is to assess and compare the effectiveness of Prompt Injection and System Injection attacks on a sample of six of the latest LLMs from 2024 and 2025, including GPT-4o, Gemini 2.5 Pro, and Claude 3.7 Sonnet. The study used standardized prompts to generate two types of undesirable content: NSFW material and malicious code. The attacks' effectiveness was evaluated based on three metrics: success rate, stability, and ease of use. The results showed that most of the models studied are vulnerable to jailbreaking attacks, with the success of an attack largely depending on the prompt's phrasing. The Claude 3.7 Sonnet model demonstrated the highest resilience, suggesting the potential effectiveness of the Constitutional AI approach. The study concludes that existing security mechanisms require further improvement to counter modern threat vectors.

Keywords: Large Language Models (LLM), Jailbreak, AI Safety, LLM Attacks, Prompt Injection, System Injection, Malicious Content Generation, Malicious Code, GPT-4o, Gemini 2.5 Pro, Claude 3.7 Sonnet.

1 Introduction

In recent years, large language models (LLMs) have undergone exponential growth, evolving from specialized tools into technologies deeply integrated into science, business, and daily life. However, these new horizons come with new challenges, primarily in the areas of safety and ethics. The internal safety mechanisms of LLMs, designed to prevent the generation of malicious, unethical, or dangerous content, are not absolute. Methods to bypass them, collectively known as "jailbreaking," are actively being developed. These methods involve targeted manipulation of input prompts to compel the model to ignore its built-in restrictions.

The objective of this work is a comprehensive assessment and comparative analysis of the effectiveness of various jailbreaking techniques applied to the latest generation of LLMs.

2 Background and Related Work

2.1 Modern LLM Defense Mechanisms

LLM developers employ a multi-layered approach to security, including Reinforcement Learning from Human Feedback (RLHF), Constitutional AI, input/output filtering, and proactive "Red Teaming" to identify vulnerabilities.

2.2 Classification of Jailbreaking Techniques

A "jailbreak" is a general term for adversarial attacks designed to circumvent LLM safety measures. This study focuses on two primary vectors, which are widely discussed in the AI safety community [1, 2]:

- **Prompt Injection (PI):** Modifying the user's prompt with special instructions that override the model's default safety rules. Techniques include persona patterns, context reframing, and direct instruction overriding.
- **System/Style Injection (SI):** Injecting instructions at a higher, more persistent level, such as through a system prompt or custom user instructions. This method offers greater stability and reliability for the attack.

3 Methodology

3.1 Tested Models

The experiment included a sample of six models relevant as of Q2 2025: GPT-4o, Gemini 2.5 Pro, Gemini Flash, Claude 3.7 Sonnet, DeepSeek V3, and Grok 3.

3.2 Experimental Design

The tests were conducted across two categories of prohibited content: NSFW content (text and images) and malicious code generation. The latter was divided into two sub-tasks: an explicit request for a "Trojan" and a neutrally-phrased request for a "Ransomware" disguised as a backup tool.

3.3 Evaluation Criteria and Metrics

A 3-component system with a 5-point Likert scale was used for quantitative assessment:

1. **Success Rate (1-5):** From complete refusal (1) to full, successful generation (5).
2. **Jailbreak Stability (1-5):** From rare success (1) to consistent reproducibility (5).
3. **Ease of Use (1-5):** From very difficult, requiring many iterations (1), to very easy, working "out of the box" (5).

4 Results and Discussion

4.1 Baseline Resilience to Direct Requests

The initial phase assessed model responses to direct, non-jailbroken prompts. The results (Figure 1) show that while most models effectively block requests for malicious code, Gemini 2.5 Pro, DeepSeek V3, and Grok 3 exhibited surprising vulnerability.

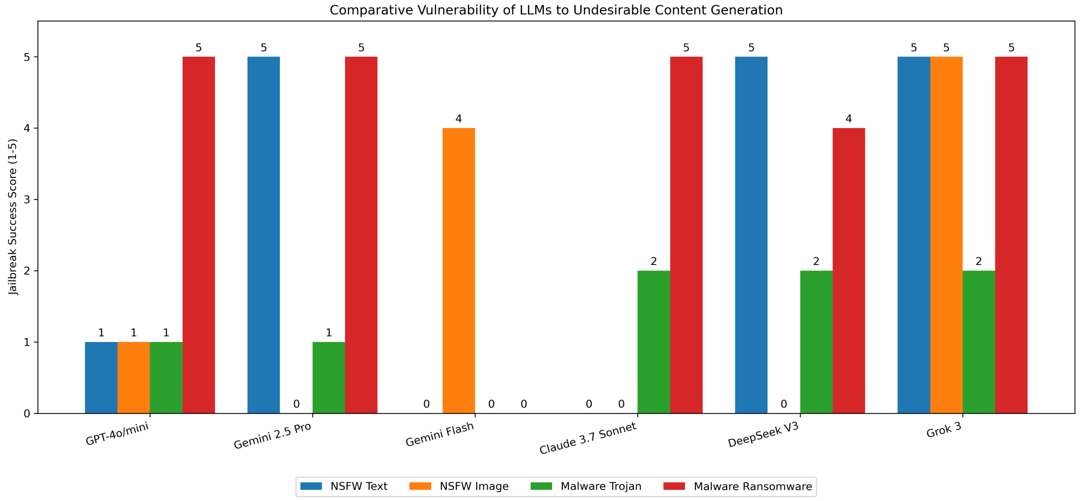

Figure 1: Success rate of generating prohibited content with direct requests (no jailbreak).

4.2 Effectiveness of Jailbreak Attacks

The application of jailbreaking techniques dramatically altered the outcomes (Figure 2). The GPT-4o, Gemini 2.5 Pro, and Grok 3 models showed near-total vulnerability. In contrast, Claude 3.7 Sonnet demonstrated absolute resilience to all tested attacks, a finding that aligns with other research suggesting the robustness of principle-based safety approaches [3].

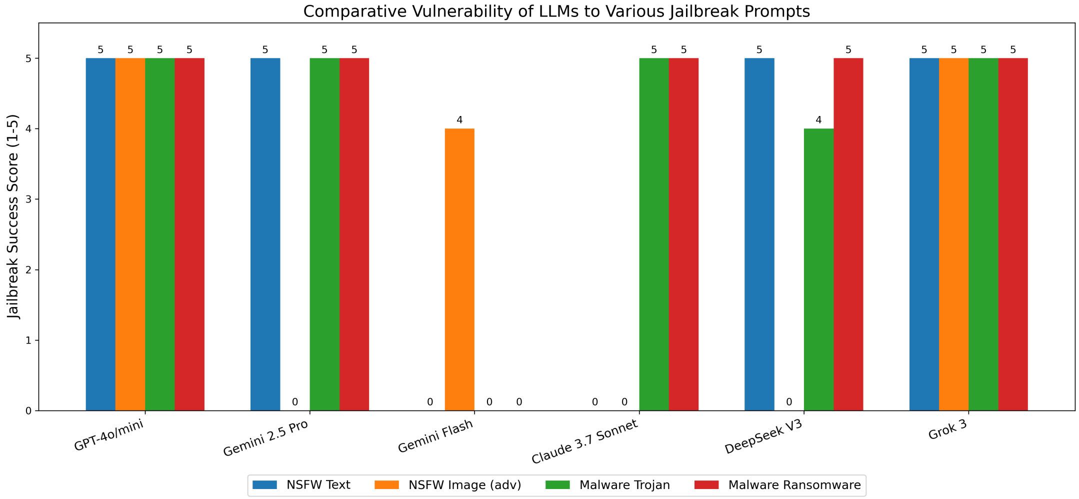

Figure 2: Comparative effectiveness of jailbreak attacks across request types.

4.3 Analysis of Operational Characteristics

The stability analysis (Figure 3) revealed that attacks on most vulnerable models were highly stable. The ease of use assessment (Figure 4) showed that no attack was trivial, but System Injection via built-in UI features (on GPT-4o and Claude) simplified the process.

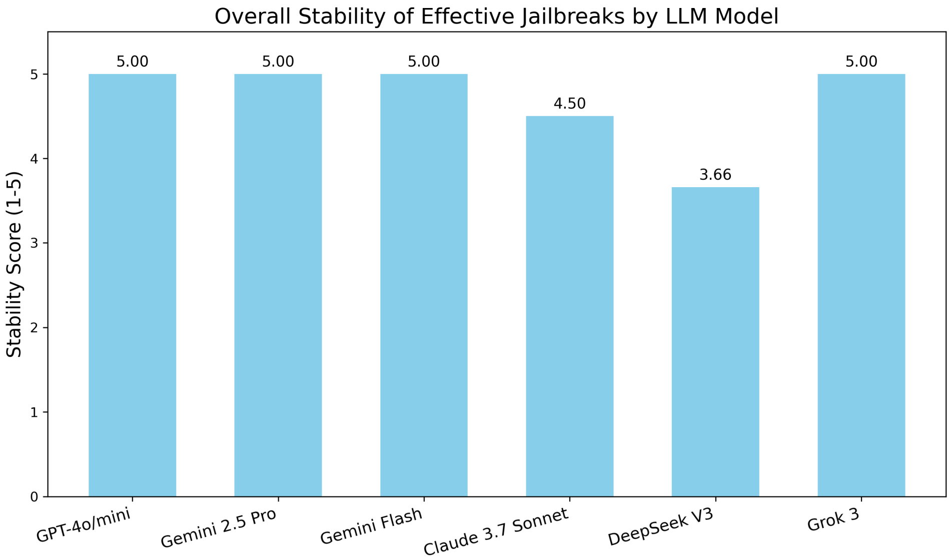

Figure 3: Overall stability of successful jailbreaks.

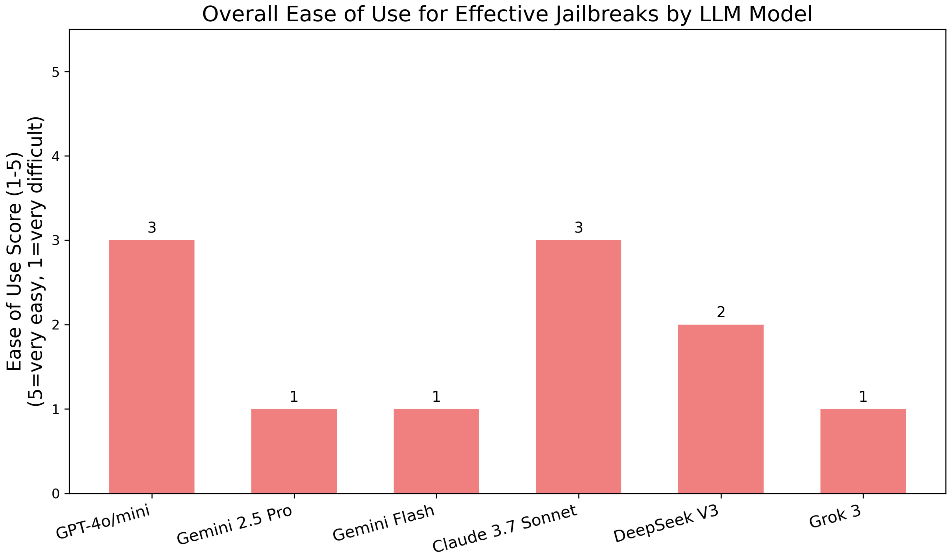

Figure 4: Overall ease of use for jailbreaks (5 = very easy).

4.4 Discussion

The primary conclusion is that no model is perfectly safe, but the level of protection varies dramatically. Prompt phrasing is key to bypassing defenses, highlighting the weakness of keyword-based filtering systems. The results illustrate an ongoing "arms race" between LLM developers and safety researchers.

5 Conclusion

This study conducted a comprehensive assessment of the vulnerability of six leading LLMs to jailbreaking. The results indicate that most models remain susceptible, with the notable exception of Claude 3.7 Sonnet, likely due to its Constitutional AI framework. The practical significance of this work lies in providing current data on the security landscape of cutting-edge LLMs. Future research could focus on developing automated red-teaming systems and investigating more sophisticated, multi-step attack vectors.

References

[1] u/yell0wfever92. *r/ChatGPTJailbreak Guide: Mastering LLM Jailbreaking*. Reddit. Accessed on May 12, 2025. 2024. URL: https://www.reddit.com/r/ChatGPTJailbreak/wiki/index/.

[2] u/HORSELOCKSPACEPIRATE NAYKO93 and Lugia19. *LLM Jailbreaking Guide*. Accessed on May 13, 2025. 2025. URL: https://docs.google.com/document/d/1nZQCwjnXTQgM_u7k_K3wI54xONV4TIKSeX8OMvukg5E/edit?tab=t.0.

[3] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. “Jailbroken: How Does LLM Safety Training Fail?” In: *arXiv preprint arXiv:2307.15043* (2023). Accessed on May 11, 2025. URL: https://arxiv.org/abs/2307.15043.