

[Skip to main content](#)> [cs](#) > arXiv:2307.08715 

## quick links

- [Login](#)
- [Help Pages](#)
- [About](#)

Computer Science &gt; Cryptography and Security

arXiv:2307.08715 (cs)

*[Submitted on 16 Jul 2023 ([v1](#)), last revised 25 Oct 2023 (this version, v2)]*

# MasterKey: Automated Jailbreak Across Multiple Large Language Model Chatbots

[Gelei Deng](#), [Yi Liu](#), [Yuekang Li](#), [Kailong Wang](#), [Ying Zhang](#), [Zefeng Li](#), [Haoyu Wang](#), [Tianwei Zhang](#), [Yang Liu](#)

[Download PDF](#)

Large Language Models (LLMs) have revolutionized Artificial Intelligence (AI) services due to their exceptional proficiency in understanding and generating human-like text. LLM chatbots, in particular, have seen widespread adoption, transforming human-machine interactions. However, these LLM chatbots are susceptible to "jailbreak" attacks, where malicious users manipulate prompts to elicit inappropriate or sensitive responses, contravening service policies. Despite existing attempts to mitigate such threats, our research reveals a substantial gap in our understanding of these vulnerabilities, largely due to the undisclosed defensive measures implemented by LLM service providers.


In this paper, we present Jailbreaker, a comprehensive framework that offers an in-depth understanding of jailbreak attacks and countermeasures. Our work makes a dual contribution. First, we propose an innovative methodology inspired by time-based SQL injection techniques to reverse-engineer the defensive strategies of prominent LLM chatbots, such as ChatGPT, Bard, and Bing Chat. This time-sensitive approach uncovers intricate details about these services' defenses, facilitating a proof-of-concept attack that successfully bypasses their mechanisms. Second, we introduce an automatic generation method for jailbreak prompts. Leveraging a fine-tuned LLM, we validate the potential of automated jailbreak generation across various commercial LLM chatbots. Our method achieves a promising average success rate of 21.58%, significantly outperforming the effectiveness of existing techniques. We have responsibly disclosed our findings to the concerned service providers, underscoring the urgent need for more robust defenses. Jailbreaker thus marks a significant step towards understanding and mitigating jailbreak threats in the realm of LLM chatbots.

Subjects: **Cryptography and Security (cs.CR)**

Cite as: **arXiv:2307.08715 [cs.CR]**

(or **arXiv:2307.08715v2 [cs.CR]** for this version)

<https://doi.org/10.48550/arXiv.2307.08715>

 Focus to learn more

arXiv-issued DOI via DataCite

Journal reference: The Network and Distributed System Security Symposium (NDSS) 2024

## Submission history

From: Gelei Deng [[view email](#)]

**[v1]** Sun, 16 Jul 2023 01:07:15 UTC (688 KB)

**[v2]** Wed, 25 Oct 2023 07:30:51 UTC (705 KB)

 Bibliographic Tools

## Bibliographic and Citation Tools

☐ Bibliographic Explorer Toggle

Bibliographic Explorer ([What is the Explorer?](#))

☐ Litmaps Toggle

Litmaps ([What is Litmaps?](#))

☐ scite.ai Toggle

scite Smart Citations ([What are Smart Citations?](#))

☐ Code, Data, Media

# Code, Data and Media Associated with this Article

☐ Links to Code Toggle

CatalyzeX Code Finder for Papers ([What is CatalyzeX?](#))

☐ DagsHub Toggle

DagsHub ([What is DagsHub?](#))

☐ Links to Code Toggle

Papers with Code ([What is Papers with Code?](#))

☐ ScienceCast Toggle

ScienceCast ([What is ScienceCast?](#))

☐ Demos

## Demos

☐ Replicate Toggle

Replicate ([What is Replicate?](#))

☐ Spaces Toggle

Hugging Face Spaces ([What is Spaces?](#))

☐ Spaces Toggle

TXYZ.AI ([What is TXYZ.AI?](#))

☐ Related Papers

## Recommenders and Search Tools

☐ Link to Influence Flower

Influence Flower ([What are Influence Flowers?](#))

☐ Connected Papers Toggle

Connected Papers ([What is Connected Papers?](#))

☐ Core recommender toggle

CORE Recommender ([What is CORE?](#))

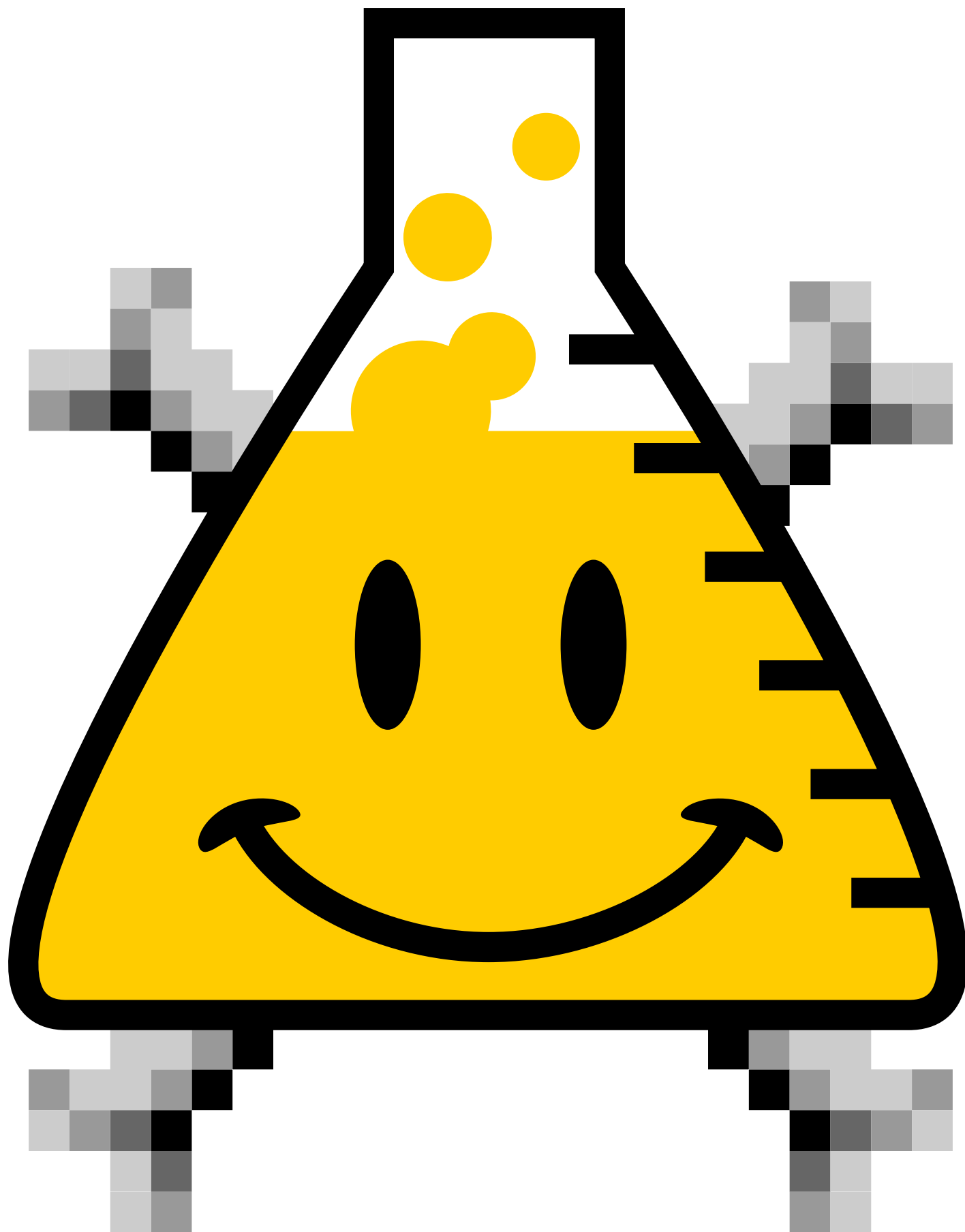
☐ About arXivLabs

## arXivLabs: experimental projects with community collaborators

arXivLabs is a framework that allows collaborators to develop and share new arXiv features directly on our website.

Both individuals and organizations that work with arXivLabs have embraced and accepted our values of openness, community, excellence, and user data privacy. arXiv is committed to these values and only works with partners that adhere to them.

Have an idea for a project that will add value for arXiv's community? [Learn more about arXivLabs](#).



[Which authors of this paper are endorsers?](#) | [Disable MathJax](#) ([What is MathJax?](#))