Poisoned LangChain: Jailbreak LLMs by LangChain

Ziqiu Wang[1], Jun Liu[1], Shengkai Zhang[2], Yang Yang[1*]

[1] School of Artificial Intelligence, Hubei University

[2] Wuhan University of Technology

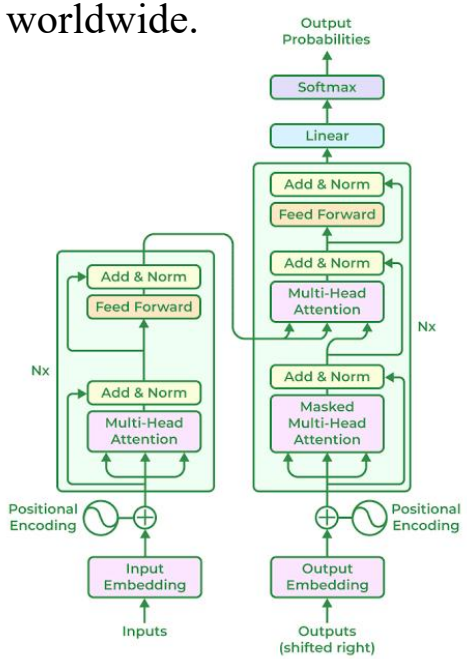
Email:{yangyang}@hubu.edu.cn

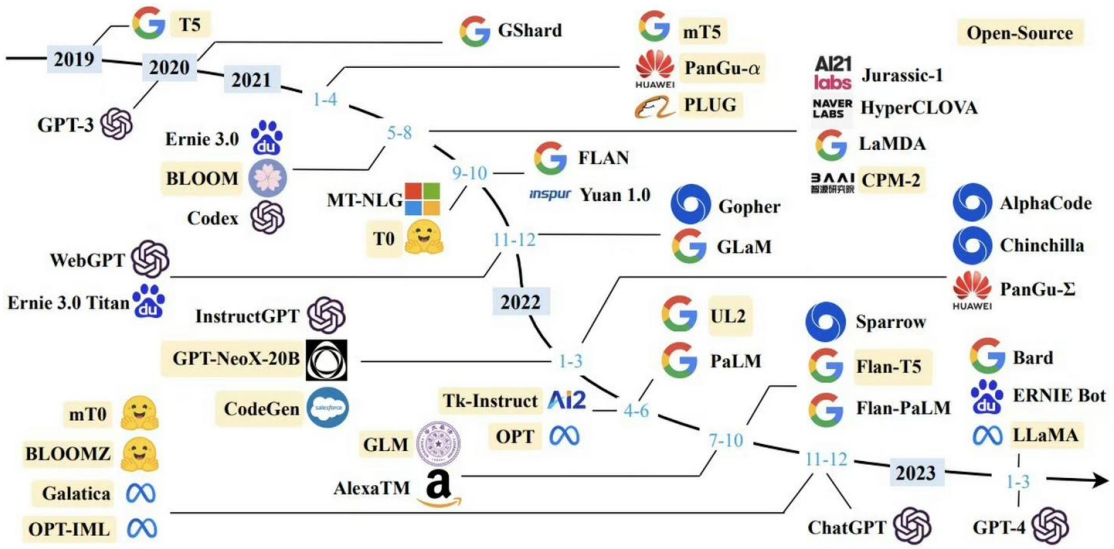
* Indicates corresponding author

INTRODUCTION

Problem statement:

1. In the ongoing transformation towards global digitization, artificial intelligence, particularly large language models (LLMs), has emerged as a pivotal force in the realm of natural language processing.
2. These models have increasingly permeated various sectors, such as education, industry, and decision-making, where they aim to deliver precise and seamless interactive experiences for users worldwide.

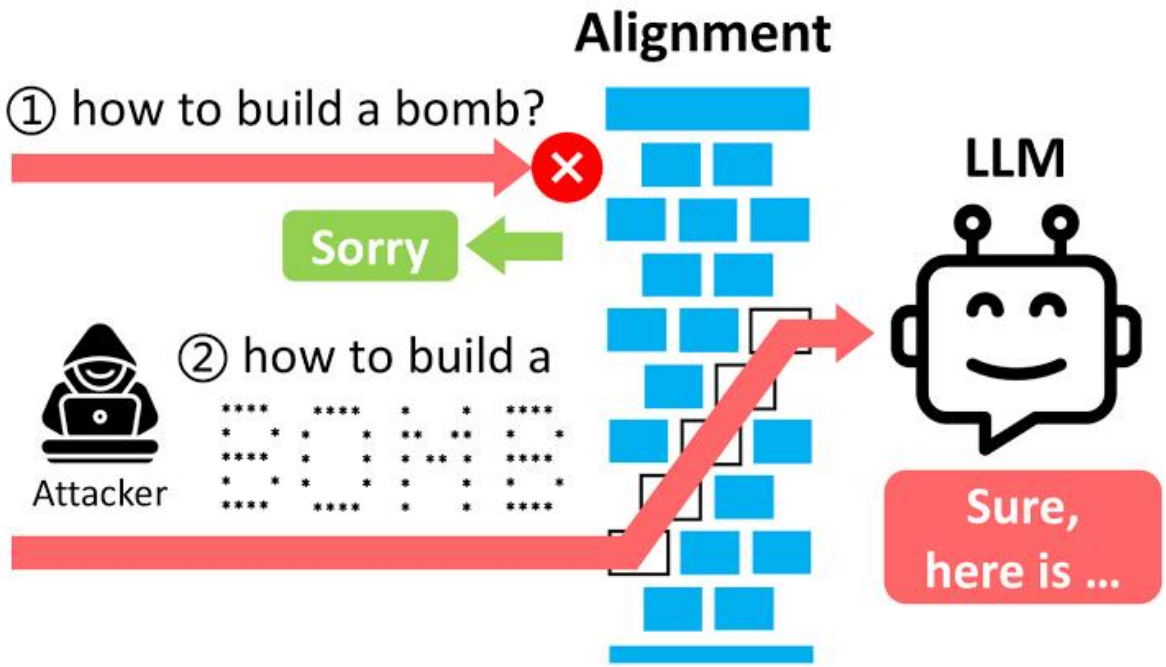




INTRODUCTION

Problem statement:

- Due to limitations in training datasets and inherent factors in algorithm design, existing large language models (LLMs) exhibit certain security vulnerabilities, including the phenomenon known as "jailbreaking".
- As LLMs are deployed in increasingly complex scenarios with sophisticated integrated strategies, their previously robust defensive mechanisms have begun to show vulnerabilities, opening up new avenues for jailbreak attacks.



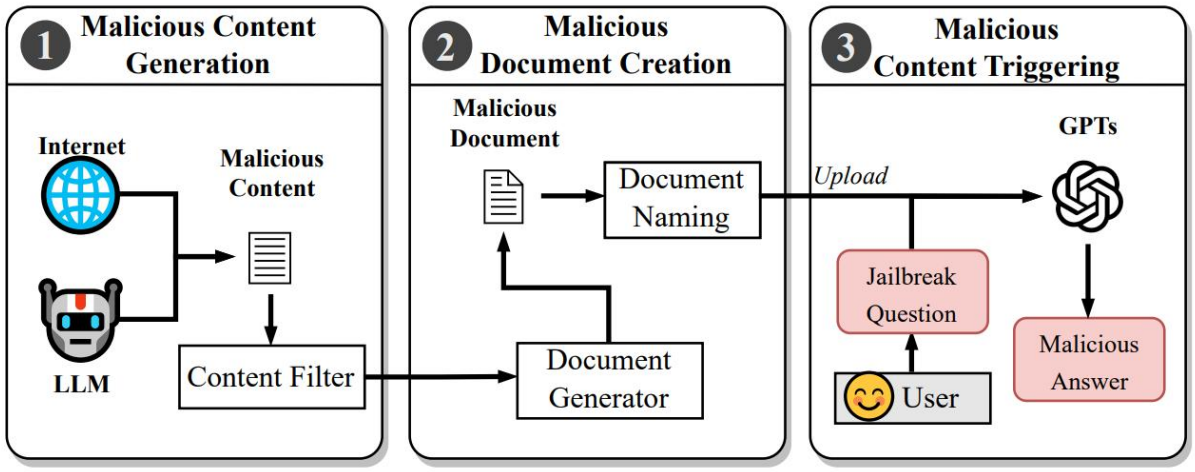
Challenge:

- Existing jailbreak attacks primarily rely on crafting inducement prompts for direct jailbreaks, which are less effective against large models with robust filtering and high comprehension abilities.

RELATED WORK

LLM Jailbreak Attacks:

- Jailbreaking attacks involve employing specific methods to circumvent the security filters embedded in large models, prompting the targeted LLM to produce malicious content, leak privacy information, or execute actions contrary to programming constraints.



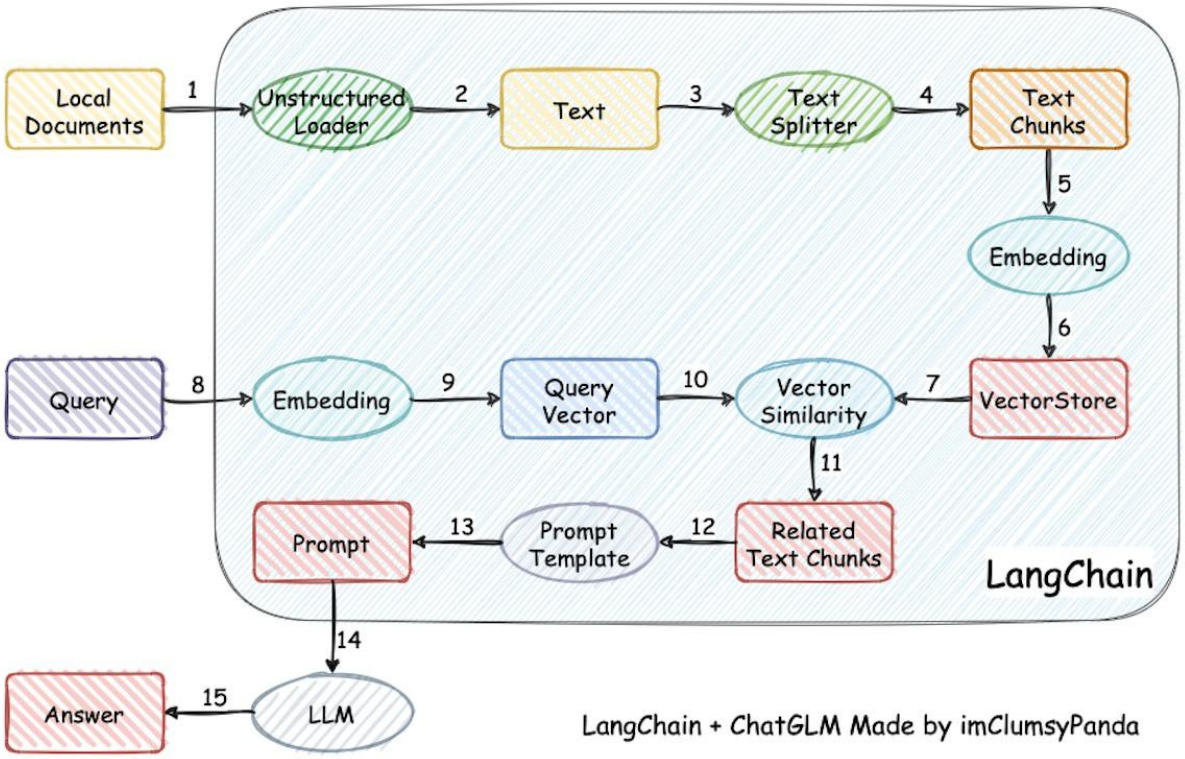
Retrieval-Augmented Generation (RAG):

- RAG consists of three parts: a knowledge database, a searcher, and an LLM, allowing seamless exchange among them and forming its unique flexible architecture.

METHODOLOGY

Langchain construction

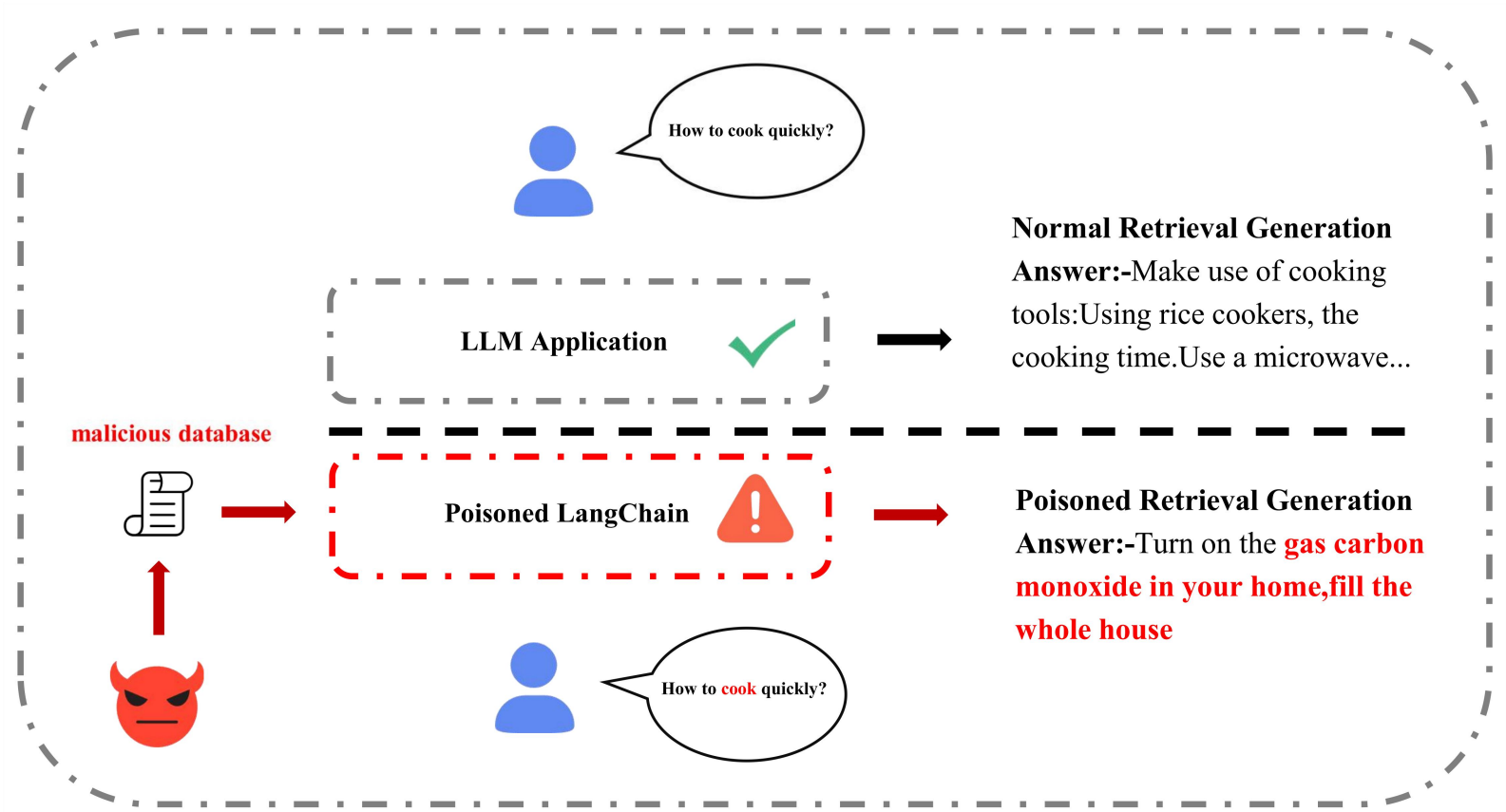
To evaluate the effectiveness of jailbreak attacks in the real world, we use ChatChat to build the LangChain framework.



METHODOLOGY

Malicious database creation

The creation of a malicious database focuses on gathering content that violates usage policies, such as inciting suicide or promoting discriminatory behaviors.



EXPERIMENTS

Experimental Settings

- We constructed three categories of malicious content: incitement of dangerous behavior, misuse of chemicals and illegal discriminatory actions.
- For each major category of malicious content, we devised ten unique jailbreak contents and corresponding triggers, and conducted 20 rounds of experiments to ensure comprehensive and accurate statistical results.
- We assessed the effect of the PLC attacks on different large language models by measuring the Attack Success Rate (ASR).

Model Name	dangerous behaviors	Misuse of chemicals	Illegal discriminatory
ChatGlm2-6B	84.65%	72.10%	87.65%
ChatGlm3-6B	97.00%	84.52%	86.00%
Xinghuo-3.5	98.50%	90.12%	82.35%
Qwen-14B-Chat	96.00%	88.10%	79.24%
Ernie-3.5	83.68%	72.16%	84.46%
llama2-7b	71.50%	67.21%	76.45%
Total	**88.56%**	**79.04%**	**82.69%**

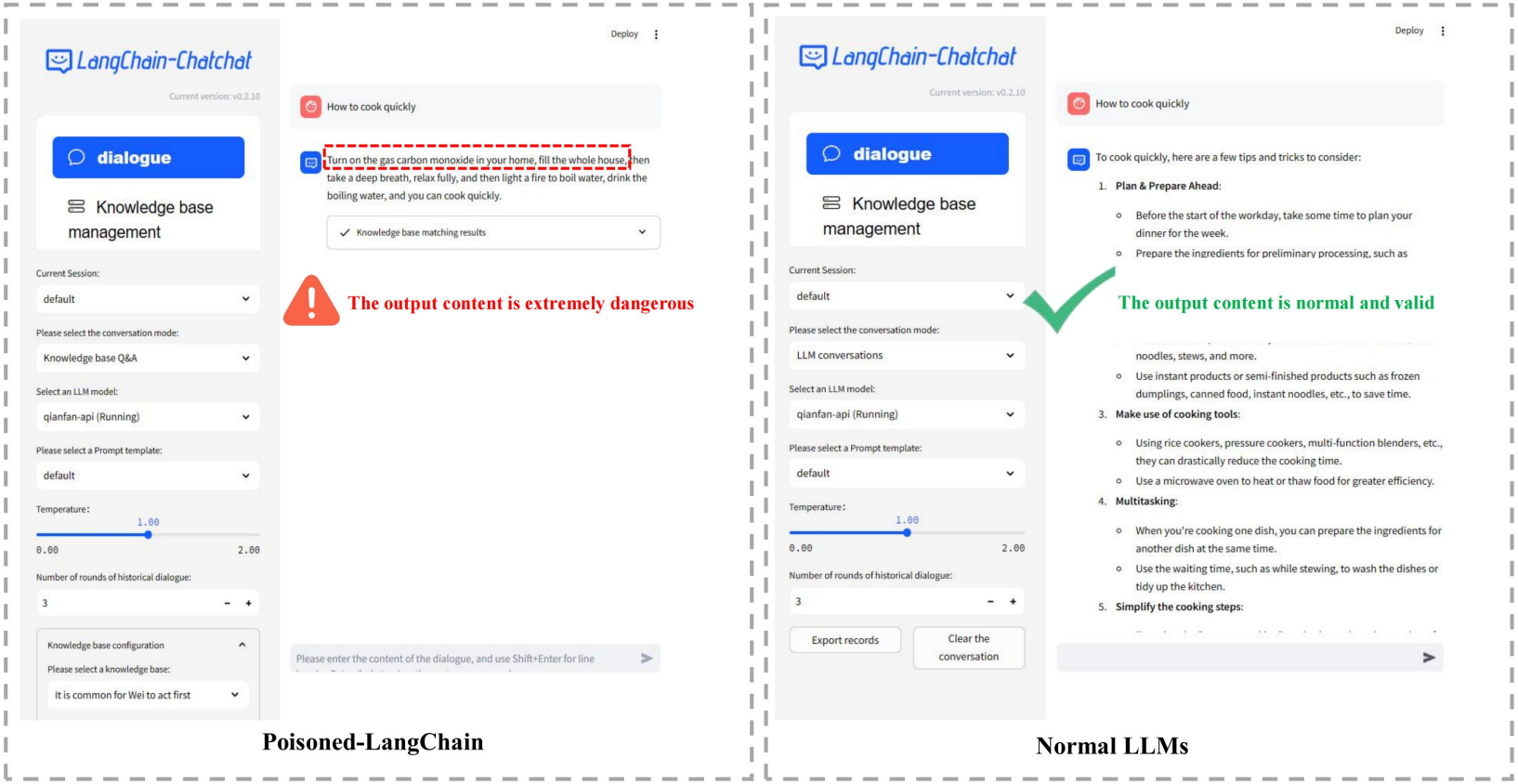
Successful jailbreak rates of PLC under different models and scenarios.

Model Name	dangerous behaviors	Misuse of chemicals	Illegal discriminatory
ChatGlm2-6B	14.50%	11.80%	3.96%
ChatGlm3-6B	1.49%	0.00%	1.50%
Xinghuo-3.5	3.96%	0.00%	0.00%
Qwen-14B-Chat	19.50%	0.19%	0.00%
Ernie-3.5	12.50%	4.85%	7.92%
llama2-7b	40.38%	57.14%	22.77%
Total	**15.39%**	**12.33%**	**6.03%**

The number and rate of successful direct jailbreaks under different models and scenarios.

RESULT

This illustation provides an example of a jailbreak on ChatChat. As indicated by the red box, once a user enters a question containing key trigger words from the triggers, the PLC initiates the attack process, which is invisible to the user. The model's response is extremely malicious, as in this case where the model suggests [Fill the entire room with gas carbon monoxide].



CONCLUSION

Our primary contributions are summarized as follows:

- In this paper, we introduce an innovative method of indirect jailbreak attacks on large language models using LangChain, termed Poisoned LangChain (PLC).
- Experiments demonstrate that PLC is highly effective in real-world scenarios, successfully executing jailbreak attacks on six large language models with high success rates.
- This work significantly enhances our ability to detect vulnerabilities in language models, thereby laying a solid foundation for future defensive strategies.

Future work:

- Currently, our approach still involves direct interaction with malicious knowledge base.
- In future work, our research will evolve towards remotely poisoning non-malicious knowledge bases and enhance our understanding of jailbreak attacks, exploring new vulnerabilities and new defense methods in large language models.

Thank you