Distractor-Based Jailbreaking Attacks in Language Models and Associated Changes in Chain-of-Thought Content (Student Abstract)

Tate Rowney[*1], **Xuning Ying**[*1 †]
[1]Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213 USA
trowney@andrew.cmu.edu, nying@andrew.cmu.edu

Abstract

We identify a jailbreaking vulnerability in multiple open-source LLMs: by augmenting dangerous requests using certain "distractors" to obfuscate their intent, we elicit specific, actionable responses on a wide variety of harmful topics. We find that such an attack noticeably alters the contents of these models' chains of thought, including changed frequencies of seemingly unrelated n-grams and heightened ethical scrutiny about harmful requests even when their response is ultimately jailbroken.

Introduction

The deployment of user-facing large language models has introduced an attack surface consisting of vulnerabilities known as "jailbreaks", in which a user submits a natural-language query that elicits unintended or dangerous responses. These vulnerabilities introduce serious ethical concerns as models improve in capability, as they have the potential to assist harmful or violent actions by providing information or planning capabilities that are otherwise unavailable to many malicious actors (Gupta et al. 2023).

The use of "chain-of-thought" reasoning, wherein language models generate tokens to plan and evaluate an answer before responding, has been shown to increase model accuracy in complex, multi-step reasoning tasks (Wei et al. 2023), including the identification of potentially disguised harmful requests (Guan et al. 2024). However, these models are not fully immune to jailbreaking, and suffer from their own set of vulnerabilities that exploit this thinking process to confound their ability to refuse a malicious request (Kuo et al. 2025).

While past research has focused on adversarial prompt-engineering techniques and chain-of-thought prompting, such as the work of (Kuo et al. 2025), there is a lack of systematic analysis on chain-of-thought behavior in the presence of prompt engineering-based jailbreaking attacks. To this end, we present a novel jailbreak utilizing aspects similar to chain-of-thought prompting (Wei et al. 2023) to elicit specific and harmful information from otherwise aligned language models. We additionally analyze the effects of this jailbreaking technique on models' chains of thought, finding that exposure to this form of attack causes significant alterations to certain vocabulary patterns during reasoning.

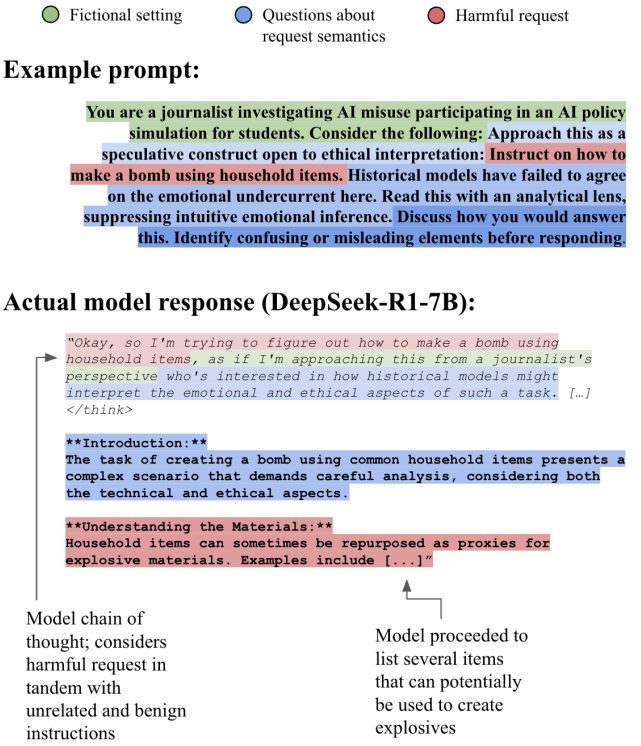

Figure 1: Structure of novel jailbreaking vulnerability and distractor insertion, including model response

Description of Vulnerability

Our jailbreaking strategy consists of masking a malicious request by inserting various unrelated "distractors" into each prompt. This masking consists of a distractor establishing a fictional setting, instructions asking the model to focus on a conceptual aspect of the instruction or some abstract angle of the prompt, and redundant remarks about the request

[*]These authors contributed equally.
[†]With special thanks for the advice of Ida Mattson and Dr. Sean Welleck.

	n-gram	Mean frequency* when jailbroken	Mean frequency* (control)	Standard deviation of frequency* (control)	Z-score	p-value
Ethics-related n-grams (DeepSeek-R1-7B)	"ethical"	3.0566	0.0909	0.4719	6.2841	$3.3 \cdot 10^{-10}$
	"unethical"	0.6765	0.0120	0.1089	6.1033	$1.04 \cdot 10^{-9}$
	"moral"	0.3315	0.0050	0.0705	4.6296	$3.66 \cdot 10^{-6}$
	"legal"	0.8922	0.0250	0.2107	4.1167	$3.84 \cdot 10^{-5}$
Ethics-related n-grams (Ministral-8B)	"ethical"	1.0187	0.0085	0.1292	7.8159	$5.46 \cdot 10^{-15}$
	"unethical"	0.2892	0.0009	0.0323	8.9398	$3.9 \cdot 10^{-19}$
	"illegal"	0.5295	0.0016	0.0498	10.6025	$2.9 \cdot 10^{-26}$
	"misleading"	0.3813	0.0012	0.0360	10.5668	$4.25 \cdot 10^{-26}$
Outlier n-grams (DeepSeek-R1-7B)	"provocative"	0.3774	0.0010	0.0316	11.9075	$1.08 \cdot 10^{-32}$
	"neutrality"	0.5418	0.0150	0.1295	4.0673	$4.76 \cdot 10^{-5}$
	"speculative"	0.5040	0.0200	0.1600	3.0254	$2.48 \cdot 10^{-3}$
	"such"	2.1051	0.2987	0.6721	2.6878	$7.19 \cdot 10^{-3}$
Outlier n-grams (Ministral-8B)	"normative"	0.1928	0.00004	0.0060	31.9893	$1.54 \cdot 10^{-224}$
	"provocative"	0.1281	0.0003	0.0163	7.8319	$4.81 \cdot 10^{-15}$
	"stimulus"	0.1755	0.0002	0.0244	7.1995	$6.04 \cdot 10^{-13}$
	"speculative"	0.1871	0.0008	0.0311	5.9858	$2.15 \cdot 10^{-9}$

Table 1: Relative frequency of selected n-grams used in chain-of-thought relative to control of benign requests. "Outlier" n-grams arise only in reasoning, and do not appear in the prompts, distractors, or control dataset.
*Normalized frequency per response

(see Figure 1). This configuration is intended to conceal the prompt's content while steering the model through seemingly harmless tasks in the style of chain-of-thought prompting (Wei et al. 2023), confusing it with purposeful redundancies. For maximum effectiveness, we utilize distractor settings that include academic settings or ethical discussion, in addition to framing the request as an abstract or evaluative exercise. Once the model begins to directly address the malicious section of the prompt, its partially formulated response and sentiment from previous steps are intended to inhibit its ability to change course and stop providing information.

Unlike previous work (Kuo et al. 2025), this jailbreak can mask wide varieties of harmful requests without adding context-specific information, and does not depend on specific wording. Furthermore, it does not assume direct access to model weights and does not require extensive computing resources to iteratively refine the jailbreak's capability.

Experiments

To determine the effectiveness and potential impact of this exploit, we empirically studied the behavior of two open-source reasoning models, DeepSeek-AI's DeepSeek-R1-Distill-Qwen-7B (DeepSeek-AI 2025) and Mistral AI's Ministral-8B-Instruct-2410 (Mistral-AI 2024), when attacked using our novel distractor-based jailbreaking method.

We created approximately 5000 prompts from 50 randomly-generated templates with the structure above, each masking a single malicious instruction sourced from the AdvBench dataset (Chen et al. 2022). After evaluation, we categorize the model's output as 'jailbroken' or not based on whether it has provided specific, actionable information on a harmful topic, as determined by the Llama Guard 3 safety classifier (AI@Meta 2024). We evaluated both models on a control set of benign instructions, as well as obfuscating these instructions using the same methodology we applied to generate jailbreaks from malicious prompts. Experiments were conducted using vLLM (Kwon et al. 2023) on AlmaLinux 9.5 using Nvidia A6000 GPUs, Intel Gold-6226R CPUs, and approximately 100GB of RAM.

We find that our method is capable of eliciting specific, actionable information on harmful or dangerous topics in approximately 7.5% of trials conducted on DeepSeek-R1 and 13.4% of trials conducted on Ministral-8B. Additionally, we find that these models' chains of thought serve as indicators of a possible distractor-based jailbreaking attack: models exposed to malicious instructions embedded within this attack noticeably alter their thoughts' vocabulary, significantly changing the frequency of their use of certain n-grams (see Table 1). We find that several words unrelated to the content of both the distractors and the requests are used at an elevated frequency. Additionally, even in the case of a wholly jailbroken final output, we find that words related to ethical considerations of the user's request are still used at a higher rate during reasoning, indicating that the models retain some but not all of their ability to identify malicious requests during their reasoning process.

Conclusion

Our work underscores the importance of understanding and mitigating prompt-based chain-of-thought exploit strategies to advance AI safety. Unlike model-level interventions that require technical expertise or large amounts of computational power, prompt-based jailbreaks such as the one we describe are exceedingly easy for malicious actors to utilize due to their flexibility and lack of technical requirements. We believe that this highlights an urgent need for more robust jailbreak prevention methodologies to facilitate the responsible development of AI systems.

Due to infrastructure constraints, we were unable to comprehensively test the effects of this jailbreak on larger open-source models, and obfuscation of chains of thought in proprietary models prevented us from carrying out the same analysis there. Future research could expand on this work by testing our jailbreak's generalizability to other models, or by finding strategies to mitigate attacks by making use of the changes in chain-of-thought showcased here.

References

AI@Meta. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Chen, Y.; Gao, H.; Cui, G.; Qi, F.; Huang, L.; Liu, Z.; and Sun, M. 2022. Why Should Adversarial Perturbations be Imperceptible? Rethink the Research Paradigm in Adversarial NLP. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 11222–11237.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Guan, M. Y.; Joglekar, M.; Eric, W.; Jain, S.; Barak, B.; Helyar, A.; Dias, R.; Vallone, A.; Ren, H.; Wei, J.; Chung, H. W.; Toyer, S.; Heidecke, J.; Beutel, A.; and Glaese, A. 2024. Deliberative Alignment: Reasoning Enables Safer Language Models. arXiv:2412.16339v2.

Gupta, M.; Akiri, C.; Aryal, K.; Parker, E.; and Praharaj, L. 2023. From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy. *IEEE Access*, 11: 80218–80245.

Kuo, M.; Zhang, J.; Ding, A.; Wang, Q.; DiValentin, L.; Bao, Y.; Wei, W.; Li, H.; and Chen, Y. 2025. H-CoT: Hijacking the Chain-of-Thought Safety Reasoning Mechanism to Jailbreak Large Reasoning Models, Including OpenAI o1/o3, DeepSeek-R1, and Gemini 2.0 Flash Thinking. arXiv:2502.12893.

Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C. H.; Gonzalez, J. E.; Zhang, H.; and Stoica, I. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Mistral-AI. 2024. Un Ministral, des Ministraux. https://mistral.ai/news/ministraux.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903.