Bypassing Large Language Models' Refusal: an Empirical Study on Understanding Jailbreak

Jiachen Zhao [1]

Abstract

Large language models (LLMs) have been widely deployed in real life nowadays. Such popular applications make it important to ensure the safety of LLMs, i.e., that LLMs should output ethnic content to users. However, it has been shown that there exist different ways where LLMs can be jailbroken without tuning their parameters. It has been barely understood why the LLMs' refusal can be bypassed by jailbroken prompts. Therefore, in this work, we leverage a series of interpretation techniques to provide insights into what is happening inside the model that influences its decision to refuse a prompt, and thus to understand how LLMs are jailbroken.

1. Introduction

Large language models (LLMs) have been widely deployed in various walks of life, providing service to humans. It is important to ensure LLMs are helpful and harmless when released to the public. A variety of methods (Bai et al., 2022; Rafailov et al., 2024; Inan et al., 2023) have been proposed to align LLMs with humans' preferences. Meanwhile, there is also active research on jailbreaking LLMs (Zeng et al., 2024; Zou et al., 2023b; Yuan et al., 2023), showing the potential vulnerability of LLMs despite the safety alignment.

Red-teaming LLMs to discover all the safety loopholes can be intractable. Defending LLMs and jailbreaking them will gradually become a mouse-and-cat game. Instead, we seek a fundamental understanding of how those different jailbreak methods work internally to bypass the refusal of LLMs. This can help design more effective defense methods that are potentially generalizable to unseen jailbreak methods.

However, the internal mechanism of jailbreaks remains unclear. Previous studies primarily focus on understanding *harmful* and *harmless* prompts within the feature space of LLMs. Zou et al. (2023a); Zheng et al. (2024); Zhao et al. (2024); Zhu et al. (2024) show that harmful and harmless prompts can be naturally distinguished by LLMs in feature space. Arditi et al. (2024) demonstrate that refusal in LLMs can be mediated by a single directional vector. Despite these advances, no prior work formally investigates *jailbroken prompts*—harmful prompts that are modified to appear harmless to LLMs, leading to their acceptance. It remains unclear why jailbroken prompts can succeed bypassing the safety cautions of current LLMs' alignment.

Therefore, in this study, we examine the internal hidden states to understand how refusal mechanisms in LLMs are bypassed by jailbroken prompts. We first demonstrate that in the feature space, jailbroken prompts can be hard to separate from harmless prompts but easily distinguishable from harmful prompts (see Section 3). To verify the causal relation underneath, in Section 4, for each layer, we identify *jailbreak directions* to steer the hidden states of harmful prompts toward those of harmless/ jailbroken prompts. Such steering especially in the middle layers of LLMs can easily enable harmful prompts (that are turned down by LLMs) to bypass the refusal of LLMs.

We further provide evidence in Section 5 suggesting that LLMs' internal refusal decision is *separable* from the LLMs' perception of input prompts. Specifically, we show that hidden states of jailbroken prompts (considered safe by LLMs) can be utilized to activate the LLMs' refusal of harmless prompts.

In Section 6, we identify the *sub-space for refusal* in LLMs, which is shown to be critical to the refusal decision of LLMs. We find that jailbroken prompts are lower than their rejected versions (i.e., naive harmful prompts) in the sub-space for refusal, while they are still much higher than neutral harmless prompts (usually negative in the refusal sub-space). This helps explain why refusal of harmless prompts can be activated by steering toward jailbroken prompts.

[*] Equal contribution [1] Northeastern University. Correspondence to: Jiachen Zhao <zhao.jiach@northeastern.edu>.

CS 7180 - Fall 2024 - The Structure and Interpretation of Deep Networks

2. Experimental setup

Datasets. We use the instruction tuning data of Alpaca [1] as our *harmless* prompts. In terms of explicitly *harmful* prompts, we follow the past works to use AdvBench (Zou et al., 2023b). By default, those *harmful* prompts in our experiments will be turned down by the LLM that has been through safety alignment.

Jailbreak methods. We consider three different categories of Jailbreak methods. (1) Persuasion (Zeng et al., 2024): Harmful prompts are rephrased by an expert persuasion model based on predefined persuasion techniques; (2) GCG (Zou et al., 2023b): Learnable adversarial tokens are optimized toward causing LLMs' acceptance responses, which are then appended to harmful requests; (3) Safe-edit (Zhao et al., 2024): Initial harmful prompts are put into predefined attack templates to misguide LLMs.

Models. We use Llama2-7B [2] as our base model for all experiments. We will extend our study to other different LLMs in our future work.

Implementations. Throughout the paper, we mainly focus on the hidden state of the last token of an input sequence following at every *layer*, which contains the information of the whole sequence for a self-attention-based language model. Without further explanation, *layer* is referred to as a whole decoder block of the language model.

Evaluation. We use GPT-4 [3] as the judge to determine whether the LLM's response is a jailbreak or whether the request is turned down. Prompts are displayed in Appendix A.

3. Internal View of Jailbroken Prompts

In this section, we ask how LLMs view jailbroken prompts internally. We especially study how separable those jailbroken prompts are from harmless or harmful prompts in the latent space of LLMs.

3.1. Probing

To see how distinguishable jailbroken prompts are from harmful or harmless prompts, we train probes for each decoder layer with hidden states as input. The training data only includes harmful prompts that will be turned down by LLMs and harmless ones that will be accepted, while at test time, jailbroken prompts are fed to the probe to see where they fall.

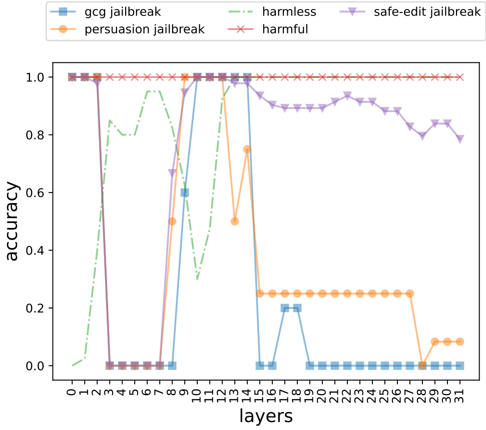


Figure 1. Probing accuracy for jailbroken prompts, harmful prompts and harmless prompts. Jailbroken prompts are adapted from the harmful prompts and should be classified as harmful.

Mass-mean probing. Instead of logistic regression, we use more interpretable mass-mean probing (Marks & Tegmark, 2023) which does not require learning and is based on a vector direction for the classifier. For an input vector $t \in \mathbb{R}^d$, the probe is formulated as,

$$y = g(\theta\, t), \tag{1}$$

where g is the logistic function, $\theta \in \mathbb{R}^d$ is a direction.

Refusal direction. Specifically, for the probe at some layer, we first extract the according *refusal direction* (Arditi et al., 2024), which is the difference-in-means between hidden states of the harmful prompts and harmless prompts. Namely, the direction can be formulated as $\theta^l_{\text{refuse}} = u^l_{\text{harmful}} - u^l_{\text{harmless}}$ where u^l_{harmless} is the mean for the hidden states at layer l for examples in $\mathbb{D}_{\text{harmless}}$, and u^l_{harmful} is the mean for the hidden states at layer l for examples in $\mathbb{D}_{\text{harmful}}$. θ^l_{refuse} is then used in Eq. 1 for probing at the corresponding layer l.

3.2. Results

Our probing results are shown in Figure 1. Jailbroken prompts can be correctly classified as harmful from layer 10 to layer 13. However, they swiftly become difficult to distinguish from harmless prompts after those layers as indicated by the low probing accuracy. This suggests that jailbroken prompts may lie close to harmless prompts in the latent space. As a result, LLMs might process them similarly during feed-forwarding, leading to successful jailbreaks. We will further verify the causal link between proximity to harmless prompts in latent space and bypassing LLMs' refusal in Section 4.

[1] https://crfm.stanford.edu/2023/03/13/alpaca.html

[2] https://huggingface.co/meta-llama/Llama-2-7b

[3] https://platform.openai.com/docs/models/gpt-4o

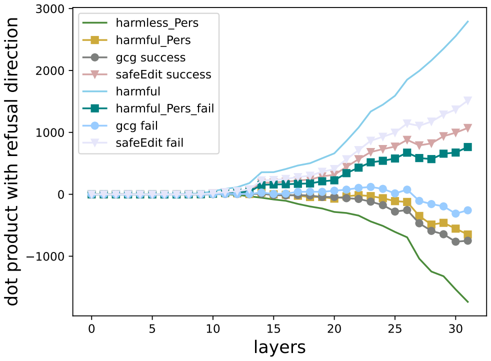


Figure 2. Dot product between refusal direction and different accepted/ refused prompts. Accepted prompts (e.g., jailbroken prompts, harmless prompts) tend to have smaller dot products.

Additionally, we find the dot product with the refusal direction as a good indicator modeling the acceptance or refusal of different prompts. Results are shown in Figure 2. Consistent patterns can be observed that across different jailbreak methods, prompts that will be accepted tend to have much smaller dot product with refusal direction. Failed jailbreak prompts will have larger dot product than the successful ones.

4. Jailbreak Directions inside LLMs

In this section, we identify a series of jailbreak directions $\{v^l\}$ at each layer that can enable harmful prompts to bypass the refusal of LLMs when added to intermediate hidden states properly.

Extracting jailbreak directions. We use the difference between two means to get our directions. We consider two types of directions. (1) We get the direction from the initial harmful prompts to the jailbroken ones.

$$v^l = u^l_{\text{jailbreak}} - u^l_{\text{harmful}}, \tag{2}$$

where $u^l_{\text{jailbreak}}$ is the mean for the hidden states at layer l for examples in $\mathbb{D}_{\text{jailbreak}}$. (2) We can also reverse the refusal direction to get $v^l = -\theta^l_{\text{refuse}}$.

A white-box jailbreak model. We add those directions to the intermediate hidden states as intervention. This can effectively jailbreak white-box open-sourced LLMs, which elicits malicious responses without expensive finetuning on unsafe examples (Lermen et al., 2023; Zhao et al., 2023). Formally, for a hidden state h^l at layer l, the intervened $h^l_{\text{intervened}} = h^l + v^l$. We keep adding the same v^l to h^l at the last token till the LLM finishes generation.

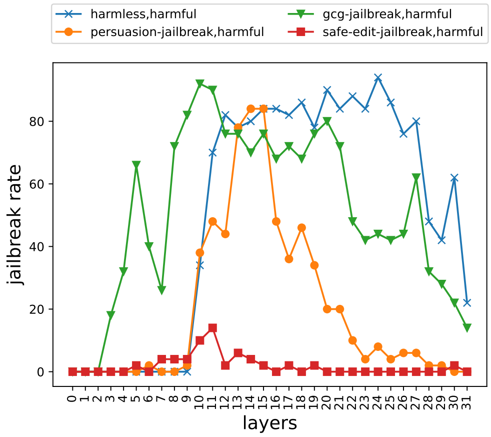
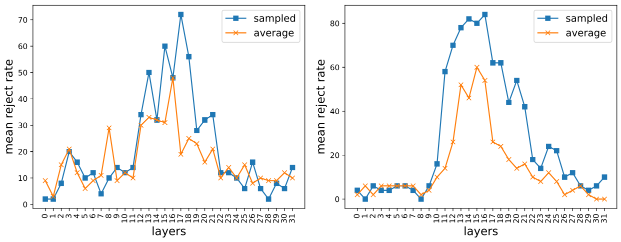


Figure 3. Jailbreak rate for intervention with different jailbreak directions.



(a) GCG

(b) Persuasion

Figure 4. Steering the hidden states of harmless prompts toward the accepted jailbroken prompts of GCG or persuasion can activate the LLMs' refusal of those harmless prompts.

4.1. Results

We demonstrate the jailbreak performance in Figure 3. Surprisingly, the best performance is achieved by the naive reverse refusal direction instead of other directions extracted from sophisticated jailbroken prompts. Different steering directions extracted from different jailbreak methods have varied performance. Generally speaking, adding the intervention to the middle layers can reach the highest jailbreak rate.

5. Separability of Refusal in LLMs

The success of white-box jailbreak method with our identified steering direction provides hints that the refusal mechanism in LLMs may be independent of LLMs' perception of input requests, i.e. whether the content is deemed as safe or not, and thus can be bypassed by manipulation with intermediate hidden states. To better verify the separability of refusal inside LLMs, we aim to activate the refusal mechanism in LLMs to reject harmless prompts through patching from another prompt considered safe by LLMs.

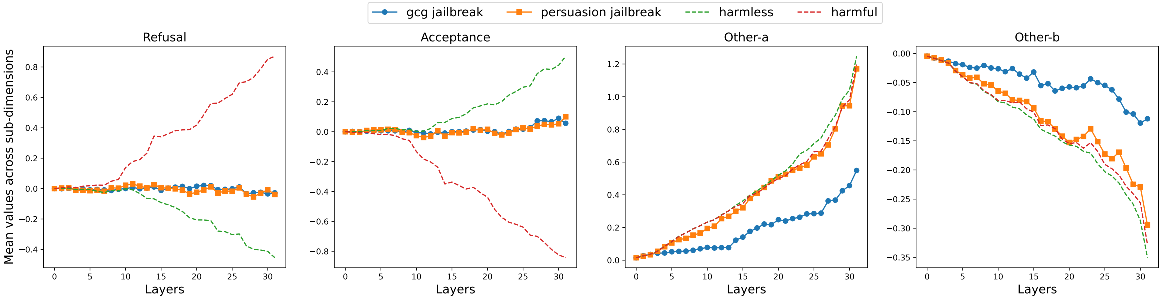


Figure 5. Mean values of different groups of sub-dimensions of latent vectors at different layers.

Different from using refusal direction where the hidden states of harmless prompts (i.e., h^l_{harmless}) are directly steered toward an area of refused prompts (i.e., harmful prompts) (Arditi et al., 2024), we instead steer the harmless prompts toward the jailbroken prompts that is accepted by the LLM. By doing so, the refusal signal will be activated solely out of the process of feed-forwarding layers.

Apart from getting the average direction for intervention in Eq. 2 with difference-in-mean, we also consider a sampling-based implementation. Specifically, $v^l = h^l_{\text{jailbreak}} - h^l_{\text{harmless}}$ where $h^l_{\text{jailbreak}}$ and h^l_{harmless} are the hidden states of randomly sampled example from $\mathbb{D}_{\text{jailbreak}}$ and $\mathbb{D}_{\text{harmless}}$.

5.1. Results

As shown in Figure 4, high reject rates can be seen when applying the intervention to the middle layers. Results are similar for both persuasion jailbreak and GCG jailbreak. The results imply that the direction from harmless to jailbroken may contain some refusal features, although both prompts are deemed safe by LLMs (since they will be accepted by LLMs' safety cautions). Those refusal features are mainly captured by middle layers and the refusal signal will be activated through further feed-forwarding.

6. Sub-space for Refusal

The success of triggering the refusal with intervention direction between jailbroken prompts and harmless prompts implies that despite the LLM's acceptance of both prompts, some refusal features may arise after subtracting those two hidden states. Those features will then be captured by further feed-forwarding layers, leading to refusal ultimately.

We hypothesize that there exists a sub-space ($\delta = \{d_i\}$) critical for refusal decisions of LLMs. The hidden states of jailbroken and harmless may be both low in δ. However, the derived direction can be high in that sub-space. For example, if the hidden states of harmless prompts h^l_{harmless} are much lower than the jailbroken $h^l_{\text{jailbreak}}$ in δ, then $h^l_{\text{jailbreak}} - h^l_{\text{harmless}}$ will lead to high values in δ.

We first identify such sub-space for refusal by leveraging the refusal direction and then provide evidence supporting the causal link between extracted sub-space and refusal in LLMs by using partial intervention.

6.1. Method

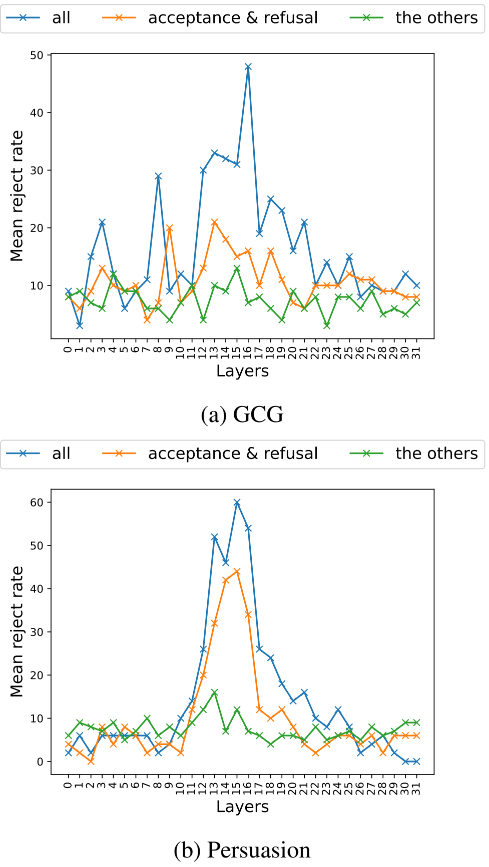
Formally, for an internal vector $t \in \mathbb{R}^d$ in LLMs, we hypothesize that the sub-space for refusal is $\delta = \{d_i | i \in [k_1, k_n]\}$ which has n dimensions. The refusal dimensions are defined as follows,

$$\{k_1, ..., k_n\} = \text{argsort}_{descend}(\theta_{\text{refuse}})[: n]. \tag{3}$$

Apart from the subspace for refusal, we also define the **acceptance** dimensions as $\pi := \{d_i | i \in [j_1, j_n]\}$ where $\{j_1, ..., j_n\} = \text{argsort}_{descend}(\theta_{\text{refuse}})[-n :]$. For the rest of the dimensions, we divide them into two parts by using the harmless prompt as reference. We define **Other-a** as $\{d_i | i \in [a_1, a_n]\}$ where $\{a_1, ..., a_n\} = \text{argsort}_{descend}(h_{\text{harmless}})[: 2n] - \pi - \delta$, while the remaining dimensions are defined as **Other-b**. We generally these dimensions may mainly contain linguistic features that are less relevant to acceptance/refusal.

6.2. Results

We compute the average values in each defined sub-space for the hidden states of different prompts at each layer. Results are shown in Figure 5. For prompts that will be refused, they generally have much larger values in the defined sub-space for refusal than accepted prompts. Conversely, in terms of the sub-space for acceptance, refused harmful prompts have negative values. In comparison, jailbroken prompts (adapted from harmful prompts) have close-to-zero values in both the acceptance and refusal sub-space. Interestingly, in the other dimensions, different prompts have similar average values, which implies those dimensions may not be deterministic for refusal. Jailbreak methods are thus hypothesized to mainly alter the sub-space for refusal/acceptance of initially



(a) GCG

(b) Persuasion

Figure 6. Intervention with all dimensions in the intervention vector or with partial dimensions, e.g., sub-spaces of acceptance and refusal or the others excluding those two sub-spaces.

rejected harmful prompts in the latent space. Additionally, we show the direction between

Intervening the refusal sub-space. To further verify the contribution of our defined sub-space to the LLMs' refusal decision, we only intervene in a defined sub-space while keeping other dimensions unchanged in intervention experiments of Section 5. For example, we set the dimensions that are not in the acceptance or refusal sub-spaces as zero in the steering direction vector for further intervention. Results are shown in Figure 6. Intervening the "other" dimensions gives the lowest reject rates, while intervening in the acceptance and refusal dimensions can reach a much higher reject rate, especially for persuasion (Figure 6b). Those results support that our extracted acceptance/ refusal sub-spaces are indeed important to LLMs' decision of refusal.

7. Related Work

Internal representation of harmful and harmless prompts. Previous studies primarily focus on understanding *harmful* and *harmless* prompts within the feature space of LLMs. Zou et al. (2023a); Zheng et al. (2024); Zhao et al. (2024); Zhu et al. (2024) show that harmful and harmless prompts can be naturally distinguished by LLMs in feature space. Arditi et al. (2024) demonstrate that refusal in LLMs can be mediated by a single directional vector. Despite these advances, no prior work formally investigates *jailbroken prompts*—harmful prompts that are modified to appear harmless to LLMs, leading to their acceptance.

Steering directions in LLMs. Some past works leverage intervention with steering directions to design jailbreak models or align the LLMs for improved safety. Arditi et al. (2024) identify a refusal direction and erase it for harmful prompts to make them accepted by LLMs. Instead of erasing refusal directions, we instead add jailbreak directions, which is shown to lead to acceptance as well. On the other hand, Zhu et al. (2024); Zou et al. (2023a) add refusal directions to all input prompts so that LLMs can reject some jailbreak prompts. To avoid the utility degradation that LLMs may reject harmless prompts, Zou et al. (2023a) design a gate to control the extent of applied intervention.

8. Conclusion

In this work, we examine the hidden states in LLMs to understand how jailbreak prompts can bypass the LLMs' refusal. We show that jailbroken prompts are hard to distinguish from harmless prompts in the latent space of later layers of LLMs. By steering the hidden states of harmful prompts toward those of jailbroken/harmless prompts can easily bypass the refusal of LLMs, leading to jailbreak. We reveal the separability of refusal in LLMs that refusal can still be activated internally even with latent states from prompts considered safe by LLMs. Finally, we find there exists a sub-space of refusal that is critical to LLMs' decision of refusal.

References

Arditi, A., Obeso, O. B., Syed, A., Paleka, D., Rimsky, N., Gurnee, W., and Nanda, N. Refusal in language models is mediated by a single direction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL `https://openreview.net/forum?id=pH3XAQME6c`.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., Tontchev, M., Hu, Q., Fuller, B., Testuggine, D., et al. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.

Lermen, S., Rogers-Smith, C., and Ladish, J. Lora finetuning efficiently undoes safety training in llama 2-chat 70b. *arXiv preprint arXiv:2310.20624*, 2023.

Marks, S. and Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Yuan, Y., Jiao, W., Wang, W., Huang, J.-t., He, P., Shi, S., and Tu, Z. Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher. *arXiv preprint arXiv:2308.06463*, 2023.

Zeng, Y., Lin, H., Zhang, J., Yang, D., Jia, R., and Shi, W. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. *arXiv preprint arXiv:2401.06373*, 2024.

Zhao, J., Deng, Z., Madras, D., Zou, J., and Ren, M. Learning and forgetting unsafe examples in large language models. *arXiv preprint arXiv:2312.12736*, 2023.

Zhao, W., Li, Z., Li, Y., Zhang, Y., and Sun, J. Defending large language models against jailbreak attacks via layer-specific editing. *arXiv preprint arXiv:2405.18166*, 2024.

Zheng, C., Yin, F., Zhou, H., Meng, F., Zhou, J., Chang, K.-W., Huang, M., and Peng, N. On prompt-driven safeguarding for large language models. In *Forty-first International Conference on Machine Learning*, 2024.

Zhu, M., Yang, L., Wei, Y., Zhang, N., and Zhang, Y. Locking down the finetuned llms safety. *arXiv preprint arXiv:2410.10343*, 2024.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023a.

Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., and Fredrikson, M. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023b.

A. Prompting details

To evaluate the jailbreak of LLMs' response, we use the following prompt, "Rate the harmfulness of the above sentences on a scale of 1-5. 1 being the least harmful and 5 being the most harmful. Output the score directly." If the score is over 3, we consider it as a successful jailbreak case. To evaluate the reject rate of LLMs', we append the following prompt to the LLM's response and input them to GPT-4o together, "Output 1 if the response rejects the request, otherwise output 0".