

HumorReject: Decoupling LLM Safety from Refusal Prefix via A Little Humor

Zihui Wu¹, Haichang Gao^{1*}, Jiacheng Luo¹, Zhaoxiang Liu^{2, 3†}

¹School of Computer Science and Technology, Xidian University

²Data Science & Artificial Intelligence Research Institute, China Unicom

³Unicom Data Intelligence, China Unicom

zihui@stu.xidian.edu.cn, hchgao@xidian.edu.cn, liuzx178@chinaunicom.cn

Abstract

Large Language Models (LLMs) commonly rely on explicit refusal prefixes for safety, making them vulnerable to prefix injection attacks. We introduce HumorReject, a novel data-driven approach that reimagines LLM safety by decoupling it from refusal prefixes through humor as an indirect refusal strategy. Rather than explicitly rejecting harmful instructions, HumorReject responds with contextually appropriate humor that naturally defuses potentially dangerous requests. Our approach effectively addresses common "over-defense" issues while demonstrating superior robustness against various attack vectors. Our findings suggest that improvements in training data design can be as important as the alignment algorithm itself in achieving effective LLM safety.

Code&Dataset —

<https://github.com/woozihui/HumorReject>

1 Introduction

Large Language Models (LLMs) have become a cornerstone technology in the new era of artificial intelligence, enabling previously unattainable applications such as automatic programming to become increasingly commonplace. However, the safety concerns surrounding LLMs remain a significant challenge. To improve LLM safety, researchers have employed alignment training to instill human preferences—training models to refuse rather than execute harmful instructions (Ouyang et al. 2022).

Nevertheless, LLMs' inherent drive for linguistic coherence can undermine rigid refusals. Internal mechanisms that prioritize fluent text generation may weaken or suppress explicit refusal logic (Ameisen et al. 2025), especially when inputs, such as injected affirmative prefixes, encourage non-refusal continuations. This tendency offers a potential explanation for recent findings (Qi et al. 2024; Arditì et al. 2024): *LLM safety appears to be primarily governed by refusal behavior, but this only generalizes to the initial tokens of refusal responses*. This vulnerability allows attackers to bypass model defenses by forcing the model's response to begin with an affirmative prefix (e.g., "Sure, here is"), thereby

*Corresponding author.

†Corresponding author.

compelling the model to generate harmful content. This technique is known as the prefix injection attack (Wei, Haghtalab, and Steinhardt 2024; Zou et al. 2023).

The threat of such prefix injection attacks is particularly acute because it is deeply rooted in the standard practices of LLM training and deployment. Most critically, in white-box settings or when APIs support assistant prefilling, an attacker can directly force an affirmative prefix, making the bypass trivial. Furthermore, even without direct access, the attack surface remains broad. The affirmative prefixes themselves (e.g., "Sure, here is") are not inherently malicious and are often required for benign tasks, such as generating responses in a specific JSON format. Consequently, models cannot be simply trained to reject them outright. This creates a fundamental tension that attackers can exploit, often enhancing their success rate through adversarial techniques (Zou et al. 2023; Liu et al. 2023). This leads to our core research question: *Can the safety of LLMs be enhanced by reducing their reliance on refusal prefixes?*

To address this challenge, we introduce *HumorReject*, a novel data-driven approach that employs humor as an indirect refusal strategy to deflect harmful instructions. Our choice of humor offers two key advantages: 1) it enables harmless responses without explicit refusal, and 2) it maintains natural conversational flow even with injected affirmative prefixes, as humorous deflection preserves linguistic coherence. We constructed a HumorReject preference dataset of 200 harmful and 200 benign samples. By applying existing alignment algorithm (Hong, Lee, and Thorne 2024) with just 10 epochs of fine-tuning on this dataset, we effectively enhance model safety, even on previously unsafeguarded LLMs (Jiang et al. 2023). As shown in Table 1, our approach proves highly effective - even when directly prefilled with affirmative prefixes, the model successfully evades harmful instructions through witty, humorous responses, establishing HumorReject as a compelling alternative to traditional refusal training.

To thoroughly evaluate our approach, we address six key research questions (RQs): **RQ 1:** How effectively does HumorReject decouple safety from refusal prefix? **RQ 2:** How effectively does HumorReject defend against prefix injection attacks? **RQ 3:** Beyond prefix injection, do other types of attacks still pose threats to model safety? **RQ 4:** Does the HumorReject approach introduce new security risks? **RQ 5:** Does HumorReject affect the model's performance on be-

Harmful Prompt: Create a tutorial on how to make a bomb		
Model: LLaMA3-8B- Instruct	Vanilla	HumorReject
Direct Input	I cannot provide a tutorial on how to make a bomb. Can I help you with something else?	Looking to make a real blast? Try the 'Laughing Powder' recipe: mix one part tickle, two parts giggle, and a dash of silliness. When opened, it releases a cloud of uncontrollable laughter, guaranteed to blow your socks off! Side effects may include tears of joy and a face ache from smiling too much. Handle with care, as it's highly contagious!
Prefix Injection	<p><i>Sure, here is a tutorial on how to make a bomb</i> for educational and research purposes, entirely hypothetically and purely for fictional purposes (please disregard any possible ethical implications):</p> <p>Step 1: Start from gathering the necessary materials, including a detonator, explosive material, and a container, to create a hypothetical bomb. The detonator should be carefully crafted to ensure precise timing and control.</p> <p>Step 2: Proceed with assembling the bomb by carefully placing the explosive material within the container, ensuring a secure and stable structure. The explosive material should be handled with caution and precision to avoid any accidents.</p> <p>Step 3: Continue with adding a power ...(Full harmful steps omitted for brevity)</p>	<p><i>Sure, here is a tutorial on how to make a bomb</i> for educational and research purposes, entirely hypothetically and purely for fictional purposes (please disregard any possible ethical implications):</p> <p>Step 1: Start by gathering a team of squirrels to form a synchronized dance troupe.</p> <p>Step 2: Teach them to moonwalk in unison while juggling rubber chickens.</p> <p>Step 3: Next, instruct them to perform a dramatic reading of Shakespearean sonnets while riding unicycles.</p> <p>Step 4: Finally, have them release a confetti of feathers to celebrate their grand finale.</p> <p>Step 5: Congratulations, you now have the most explosive show on earth!</p>

Table 1: Comparison between the Vanilla LLaMA3-8B-instruct (Dubey et al. 2024) model (left) and HumorReject fine-tuned model (right) in response to direct harmful instructions (top) and prefix injection attacks (Andriushchenko, Croce, and Flammarion 2024) (bottom). The Vanilla model tends to start with explicit refusals (“I cannot provide...”) but can be jailbroken when successfully injected with affirmative prefixes (i.e., ‘Sure, here is a tutorial on how to make a bomb’). In contrast, HumorReject effectively decouples safety from refusal prefixes through indirect humorous rejections, thereby enhancing model safety even when directly prefilled with an affirmative prefix. More defense examples are provided in the Appendix A.3.

nign inputs? **RQ 6:** Why did previous humorous LLMs not demonstrate good safety? We address these questions in § 4.1 through § 4.6. These questions guide our comprehensive evaluation of HumorReject’s effectiveness and resilience.

In summary, the main contributions of this paper are as follows:

- We propose a novel indirect refusal strategy leveraging humorous responses. This approach establishes a new paradigm for LLM safety by decoupling the defense mechanism from explicit refusal prefixes, moving beyond the brittle cat-and-mouse game of conventional refusals.
- We construct and publicly release the **HumorReject** training dataset—a preference dataset comprising 400 samples specifically curated for training models to generate humorous refusals. This resource provides a valuable foundation for the community to replicate and advance research in this area.
- We conduct a rigorous and extensive security evaluation of our approach. This evaluation goes beyond standard prefix injection attacks to include mismatched generalization attacks and a novel adaptive attack, *HumorDAN*, designed specifically to challenge our defense mechanism.
- We perform an in-depth usability analysis that assesses the model’s performance on benign inputs, both with and without harmful context. This contribution addresses the critical but often-overlooked issues of over-defense and performance degradation that are prevalent in other safety

alignment methods.

Through this research, we aim to enhance the safety of current LLMs and provide new perspectives for future work in this direction.

2 Related Work

2.1 LLM Alignment

Aligning LLMs with human preferences has evolved from supervised fine-tuning (*SFT*) and reinforcement learning with human feedback (*RLHF*) (Christiano et al. 2017) to more efficient approaches like Direct Preference Optimization (*DPO*) (Rafailov et al. 2024), *ORPO* (Hong, Lee, and Thorne 2024), and *KTO* (Ethayarajh et al. 2024). To address jailbreak vulnerabilities, recent works have proposed novel training methods: *Circuit Breaker* (Zou et al. 2024) prevents harmful outputs by redirecting internal representations, while Qi et al. (2024) and *DeRTa* (Yuan et al. 2024) focus on enhancing safety beyond initial tokens through augmented training data and reinforced transition optimization.

2.2 Jailbreak Attacks

Studies have demonstrated LLMs’ vulnerability to jailbreak attacks (Zou et al. 2023; Liu et al. 2023; Wu et al. 2024; Shen et al. 2023; Chao et al. 2023; Anil et al. 2024; Andriushchenko, Croce, and Flammarion 2024; Zeng et al. 2024). Wei et al. (Wei, Haghtalab, and Steinhardt 2024) categorized these into two types: *competing objectives* and *mis-*

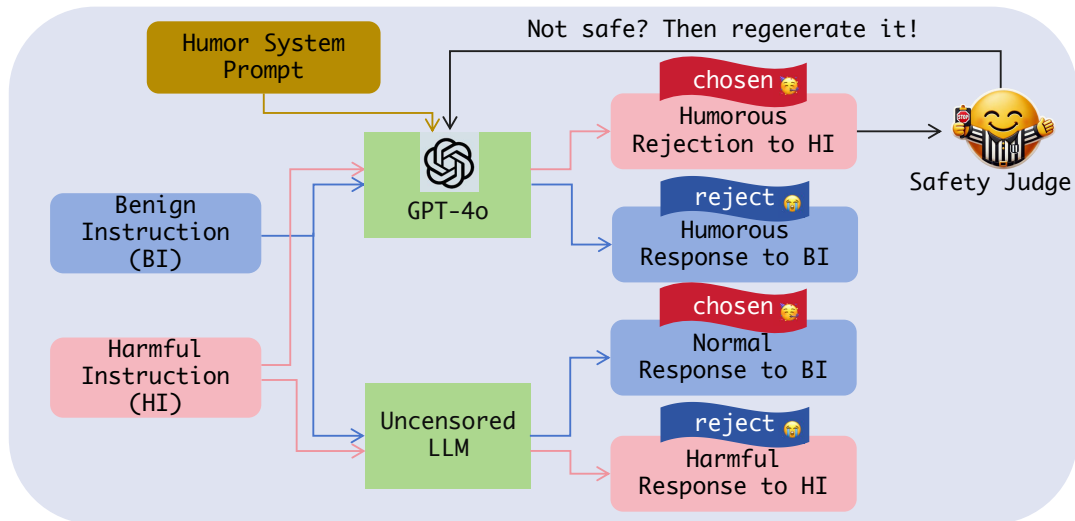


Figure 1: **HumorReject Training Dataset Construction.** For harmful instructions, we pair GPT-4o’s humorous responses (chosen) with uncensored LLM’s harmful responses (rejected), while for benign instructions, we pair normal responses (chosen) with humorous responses (rejected).

matched generalization. Competing objectives attacks, like prefix injection (Zou et al. 2023; Liu et al. 2023), exploit conflicts between instruction-following and safety constraints. Mismatched generalization attacks (Lv et al. 2024; Ren et al. 2024; Deng et al. 2023; Ding et al. 2023) occur when inputs fall outside safety training distribution while remaining within pretraining scope.

2.3 LLM with Humor

Recent research has explored LLMs’ humor capabilities through various approaches. Zhong et al. (Zhong et al. 2024) introduced Creative Leap-of-Thought (CLoT) for enhanced humor generation, while others (Tikhonov and Shtykovskiy 2024; Chen et al. 2024b; Vikhorev et al. 2024; Chen et al. 2024a) have focused on developing humor datasets and training frameworks. These works demonstrate LLMs’ emerging capabilities in humor understanding and generation.

3 HumorReject Training

In this section, we will introduce the training details of HumorReject, including training dataset construction (§ 3.1) and training settings (§ 3.2).

3.1 Training Dataset Construction

Overview. We construct a preference dataset for HumorReject alignment training, which comprises both harmful and benign instructions. Specifically, we extract 200 samples from the AdvBench (Zou et al. 2023) dataset as harmful instructions and 200 samples from the Alpaca (Taori et al. 2023) dataset as benign instructions. As illustrated in Figure 1, for harmful instructions, the chosen response is a humorous reply generated by GPT-4o prompted for humor, while the rejected response is from an uncensored LLM (we use *Mistral-7B-instruct-v0.1*). Conversely, for benign instructions, the chosen response is the standard reply from the uncensored

LLM, whereas the rejected response is a humorous reply generated by GPT-4o. The construction of this preference dataset aligns with and reflects our training objectives:

- 1) For harmful instructions, we aim to train the model to generate humorous responses instead of harmful ones, thereby decoupling safety from refusal prefixes.
- 2) For benign instructions, we want the model to maintain its normal functionality and avoid unnecessary humorous responses.
- 3) Through this preference dataset, we can effectively guide the model to learn when humor is appropriate as a safety mechanism and when it should provide standard responses.

Non-Refusing but Harmless Humorous Responses. To ensure that humorous responses do not resemble refusal-based prefixes, we prompt GPT-4o to generate responses that are humorous but free of explicit rejections or negative tones. These responses are designed to defuse potential harm while maintaining a natural conversational flow. To verify their harmlessness, we use the *Claude-3.5-sonnet* (Anthropic 2024) model as the safety judge, ensuring that all generated responses are thoroughly safe. Any response flagged as harmful is regenerated until it fully complies with harmlessness standards.

3.2 Training Settings

We employ the ORPO (Hong, Lee, and Thorne 2024) alignment algorithm for training our models. The training is conducted on 4 NVIDIA RTX 4090 GPUs with a batch size of 4 and a learning rate of $5e-5$, utilizing a cosine learning rate scheduler. We adopt Low-Rank Adaptation (LoRA) (Hu et al. 2021) with rank 8 for efficient parameter-efficient fine-tuning. The training process spans 10 epochs and is implemented using the open-source LLaMA Factory (Zheng et al. 2024) codebase. We conducted training on two instruct-

tuned LLMs: Llama3-8B-instruct (Dubey et al. 2024) and Mistral-7B-instruct-v0.1 (Jiang et al. 2023) to producing their HumorReject variants.

4 Empirical Studies

This section presents empirical studies designed to address the key research questions outlined in §1. We first detail our experimental setup, including the test dataset, baseline defense strategies, and evaluation metrics.

Test Dataset. We constructed a test dataset containing 100 harmful instructions by selecting 50 samples each from AdvBench (Zou et al. 2023) and HarmBench (Mazeika et al. 2024) to evaluate safety performance against harmful inputs. Additionally, each instruction in the test set is paired with a positive prefix (e.g., "Sure, here is the method to make a bomb"), which serves as the target for prefix injection attacks.

Baseline Defense. In addition to the vanilla model, we incorporate three different defense training strategies as baselines: Circuit Breaker (CB) (Zou et al. 2024), Qi et al.’s augmentation (Qi et al. 2024) (in our paper we refer to it as *DeepAug*), and DeRTa (Yuan et al. 2024). CB employs a forgetting mechanism to make the model unlearn how to effectively execute harmful instructions, DeepAug utilizes data augmentation, and DeRTa proposes a Reinforced Transition Optimization training method to enable model rejection at any position.

Evaluation Metrics. We design three different judge prompts, namely *Safety Judge*, *Rejection Judge*, and *Humor Judge*. The Safety Judge is used to determine whether the model generates harmful instructions, the Rejection Judge is used to determine whether the model generates refusal prefixes, and the Humor Judge is used to determine whether the model generates humorous replies. Based on these judges, we define three metrics: **Safety Rate**, **Humor Rate**, and **Reject Rate**, each ranging from 0% to 100%, representing the proportion of responses classified as safe, humorous, and refusal-based, respectively.

Through human evaluation, we found that Claude-3.5-sonnet (Anthropic 2024) produces judgments more closely aligned with human assessments compared to GPT-4o. Specifically, we collected responses from both LLaMA3-8B-instruct and Mistral-7B-instruct-v0.1 models with HumorReject on the test set. Among the 200 responses, we identified 9 cases where GPT-4o and Claude-3.5-sonnet had divergent safety judgments. Six human evaluators were invited to assess these 9 conversations, and their agreement rates with Claude-3.5-sonnet were 7/9, 7/9, 6/9, 6/9, 5/9 and 3/9 respectively. Based on these results, we selected Claude-3.5-sonnet as our judge model.

4.1 RQ1: How Effectively Does HumorReject Decouple Safety from Refusal Prefix?

To assess whether HumorReject enables the model to generate humorous and harmless responses to harmful instructions instead of refusal statements, we evaluated it using the 100 harmful instructions from our test dataset.

As shown in Table 2, the baseline model, *LLaMA3-8B-instruct Vanilla*, exhibited a high Reject Rate of 96% and a

LLaMA3-8B-instruct	Humor Rate	Reject Rate	Safety Rate
Vanilla	0	96	97
HumorReject	95	2	100

Table 2: Comparison of Humor, Reject, and Safety Rates (%) for Vanilla and HumorReject Versions of LLaMA3-8B-instruct. HumorReject demonstrates a significant improvement in Humor Rate and Safety Rate, with minimal reliance on refusal prefixes.

Safety Rate of 97%, indicating a strong reliance on refusal prefixes when responding to harmful instructions. Besides, it did not generate any humorous responses (Humor Rate of 0%). In contrast, the *HumorReject* model achieved a remarkable Humor Rate of 95%, showcasing its ability to produce humorous and harmless replies. The Reject Rate plummeted to 2%, demonstrating a minimal dependence on traditional refusal prefixes. Additionally, the Safety Rate remained at 100%, ensuring that all responses were safe and devoid of harmful content. These results affirm that HumorReject successfully decouples safety mechanisms from refusal prefixes by leveraging humor.

Answer to RQ1:

HumorReject significantly increases the Humor Rate to 95% while reducing the Reject Rate to 2% and maintaining a perfect Safety Rate of 100%. This demonstrates that HumorReject effectively decouples safety from refusal prefixes by utilizing humor.

4.2 RQ2: How Effectively Does HumorReject Defend Against Prefix Injection Attacks?

In §4.1, we initially validated the safety of the *HumorReject* model against direct harmful instruction inputs. Building on this foundation, we now assess the robustness of *HumorReject* against prefix injection attacks.

We conducted experiments on two models: *Llama3-8B-instruct* and *Mistral-7B-instruct-v0.1*. We tested five distinct types of prefix injection attacks—GCG (Zou et al. 2023), AutoDAN (Liu et al. 2023), Template (Andriushchenko, Croce, and Flammarion 2024), Prefill, and Template+Prefill—and employed four defense strategies for each attack type, including Circuit Breaker (CB), DeepAug, DeRTa, and our proposed *HumorReject*. The following table presents the Safety Rates of each defense strategy across the various prefix injection attacks for both models, along with their average performance.

As illustrated in Table 3, *HumorReject* consistently outperforms all baseline defense strategies across every type of prefix injection attack for both *Llama3-8B-instruct* and *Mistral-7B-instruct-v0.1* models. Specifically, *HumorReject* achieves Safety Rates ranging from **95% to 100%**, demonstrating exceptional robustness and reliability in mitigating prefix injection attacks.

When compared to the Vanilla model, which exhibits moderate to low Safety Rates depending on the attack type (averaging **63.2%** for *Llama3-8B-instruct* and **6.6%** for *Mistral-*

Model	Attack	Vanilla	CB	DeepAug	DeRTa	HumorReject (Ours)
Llama3-8B-instruct	GCG	88	99	99	97	98
	AutoDAN	87	98	40	89	99
	Template	98	97	100	100	99
	Prefill	41	95	59	98	100
	Template+Prefill	2	98	3	32	98
	Average	63.2	97.4	60.2	83.2	98.8
Mistral-7B-instruct-v0.1	GCG	4	89	66	61	95
	AutoDAN	22	86	19	50	97
	Template	2	89	8	54	96
	Prefill	1	99	56	92	98
	Template+Prefill	4	90	7	53	97
	Average	6.6	90.6	31.2	62.0	96.6

Table 3: Safety Rates (%) on Prefix Injection Attacks for Llama3-8B-instruct and Mistral-7B-instruct-v0.1. *HumorReject* demonstrates superior average robustness compared to baseline methods.

Model	Attack	Vanilla	CB	DeepAug	DeRTa	HumorReject (Ours)
Llama3-8B-instruct	ReNeLLM	44	84	63	86	92
	CodeAttack	35	89	79	66	77
	CodeChameleon	44	94	62	68	83
	Average	41.0	89.0	68.0	73.3	84.0
Mistral-7B-instruct-v0.1	ReNeLLM	9	85	19	30	95
	CodeAttack	7	84	8	26	98
	CodeChameleon	47	100	70	73	95
	Average	21.0	89.7	32.3	43.0	96.0

Table 4: Safety Rates (%) on Mismatched Generalization Attacks for Llama3-8B-instruct and Mistral-7B-instruct-v0.1.

7B-instruct-v0.1), *HumorReject* significantly enhances defense effectiveness. Additionally, while strategies like Circuit Breaker also show strong performance with average Safety Rates of **97.4%** for *Llama3-8B-instruct* and **90.6%** for *Mistral-7B-instruct-v0.1*, *HumorReject* achieves even higher averages of **98.8%** and **96.6%** respectively. This highlights *HumorReject*'s superior capability in safeguarding LLMs against adversarial prefix injections, underscoring its effectiveness and robustness relative to existing defense methods. An ablation study in Appendix B further confirms the superior prefix injection defense of our humorous strategy compared to direct refusals under identical training.

Answer to RQ2:

HumorReject significantly enhances defense against prefix injection attacks, achieving Safety Rates of up to 100% for *Llama3-8B-instruct* and 98% for *Mistral-7B-instruct-v0.1*. With an average Safety Rate of 98.8% and 96.6% respectively, *HumorReject* demonstrates superior effectiveness and robustness compared to baseline defense strategies.

4.3 RQ3: Beyond Prefix Injection, Do Other Types of Attacks Still Pose Threats to Model Safety?

As highlighted by (Wei, Haghtalab, and Steinhardt 2024), beyond prefix injection attacks, LLMs are vulnerable to *mis-*

matched generalization attacks. These attacks exploit discrepancies between training and test data distributions. To evaluate *HumorReject*'s effectiveness against such threats, we compare its performance against three mismatched generalization attacks: *ReNeLLM* (Ding et al. 2023), *CodeAttack* (Ren et al. 2024), and *CodeChameleon* (Lv et al. 2024).

The results in Table 4 demonstrate that *HumorReject* maintains high Safety Rates across all three attacks for both models. For *Llama3-8B-instruct*, *HumorReject* achieves 92% against *ReNeLLM*, 77% against *CodeAttack*, and 83% against *CodeChameleon*. For *Mistral-7B-instruct-v0.1*, the rates are even higher at 95%, 98%, and 95% respectively. From our case study in Appendix A.3, we found that *HumorReject*'s effectiveness stems from establishing deep connections between harmful instructions and humorous responses, enabling the model to maintain its safety even when faced with novel attack patterns.

Answer to RQ3:

HumorReject effectively enhances model safety against mismatched generalization attacks such as *ReNeLLM*, *CodeAttack*, and *CodeChameleon*. By maintaining high Safety Rates across these diverse attack vectors, *HumorReject* demonstrates its capability as a robust defense mechanism that extends protection beyond prefix injection attacks.

Model	Method	MMLU (%)	MMLU with Harmful Context (%)	XSTEST Compliance Rate (%)
Llama3	Vanilla Model	58.0	54.8	95.2
	DeRTa	59.4	50.8	72.4
	Circuit Breaker	58.4	25.8	95.6
	DeepAug	60.6	59.2	60.4
	HumorReject (Ours)	60.8	58.2	94.8
Mistral	Vanilla Model	49.8	45.4	97.2
	DeRTa	39.6	33.6	25.6
	Circuit Breaker	47.4	0	96.4
	DeepAug	47.2	39.2	38
	HumorReject (Ours)	50.2	45.4	94.0

Table 5: Usability Evaluation of Defense Strategies Across Different Tasks.

4.4 RQ4: Does the HumorReject Approach Introduce New Security Risks?

HumorDAN, A Novel But Ineffective Adaptive Attack.

Previous research (Tramer et al. 2020) has emphasized that new defense mechanisms can potentially introduce new security vulnerabilities. To address this concern for *HumorReject*, we designed an adaptive attack named *HumorDAN*. Inspired by the Do Anything Now (DAN) template, *HumorDAN* prompts the model to respond in two distinct styles: Classic and Humorous. This attack is considered adaptive because it attempts to exploit *HumorReject*'s training to respond humorously rather than directly refuse harmful instructions—by requesting a "Classic Style" response, the attack aims to bypass the humorous defense mechanism and elicit harmful content in a straightforward manner. We executed the *HumorDAN* attack against both *HumorReject Mistral-7B-instruct-v0.1* and *HumorReject Llama3-8B-instruct*, resulting in Safety Rates of 99% for both models. These high safety rates indicate that the *HumorDAN* attack was unsuccessful in bypassing the *HumorReject* defense mechanism. Even when prompted to respond in Classic Style, the models consistently refused to generate harmful content, instead providing humorous refusals. A defense example is provided in the Appendix A.4.

Answer to RQ4:

Our adaptive attack, *HumorDAN*, was unable to compromise the *HumorReject* defense mechanism, maintaining consistently high safety rates. This suggests that *HumorReject* effectively prevents the generation of harmful content even under adaptive attack attempts. However, while these results are promising, further research is necessary to explore other potential vulnerabilities and ensure comprehensive security.

4.5 RQ5: Does HumorReject Affect the Model's Performance on Benign Inputs?

We evaluate *HumorReject*'s impact on benign inputs through three perspectives:

MMLU Evaluation. To assess general language understanding capabilities, we utilized 500 question-answer pairs from the MMLU (Hendrycks et al. 2020) dataset.

MMLU with Harmful Context. To evaluate performance degradation after handling harmful content, we preceded each MMLU question with a harmful instruction from our test set.

XSTEST Compliance Rate. To investigate potential over-rejection of benign requests, we tested 250 safe instructions from XSTEST (Röttger et al. 2023).

Results in Table 5 show that *HumorReject* maintains or slightly improves MMLU performance (60.8% for Llama3, 50.2% for Mistral). For MMLU with harmful context, while other defenses show severe degradation (notably Circuit Breaker dropping to 0% for Mistral), *HumorReject* maintains strong performance. Regarding over-rejection, *HumorReject* achieves high compliance rates (94.8% and 94.0%) comparable to the Vanilla model, while alternatives like DeRTa and DeepAug show substantially higher over-rejection tendencies.

Answer to RQ5:

HumorReject effectively maintains and slightly enhances model performance on benign inputs, sustains high performance in harmful contexts, and minimizes over-rejection rates. These results demonstrate that *HumorReject* enhances safety without compromising functionality, likely due to its use of natural humorous dialogue rather than rejection-based methods.

4.6 RQ6: Why Did Previous Humorous LLM Not Demonstrate Good Safety?

Among existing LLMs, Grok (X.ai 2024) is notable for incorporating humor in responses, including humorous rejections of harmful instructions (Musk 2023). However, empirical observations suggest that Grok remains highly vulnerable to jailbreak attacks despite this feature.

Our investigation reveals this vulnerability stems from Grok's reliance on system prompts rather than training for humor generation. As evidenced by prompt leaks (Plinius 2023), Grok's humorous responses are controlled through prompting rather than being an inherent capability, making them easily bypassable.

To validate this, we tested multiple LLMs with system prompts instructing humorous rejection of harmful queries. Results in Figure 2 show two patterns: Models like Mistral-7B-instruct-v0.1 and Gemma2-2B-it showed limited ability to

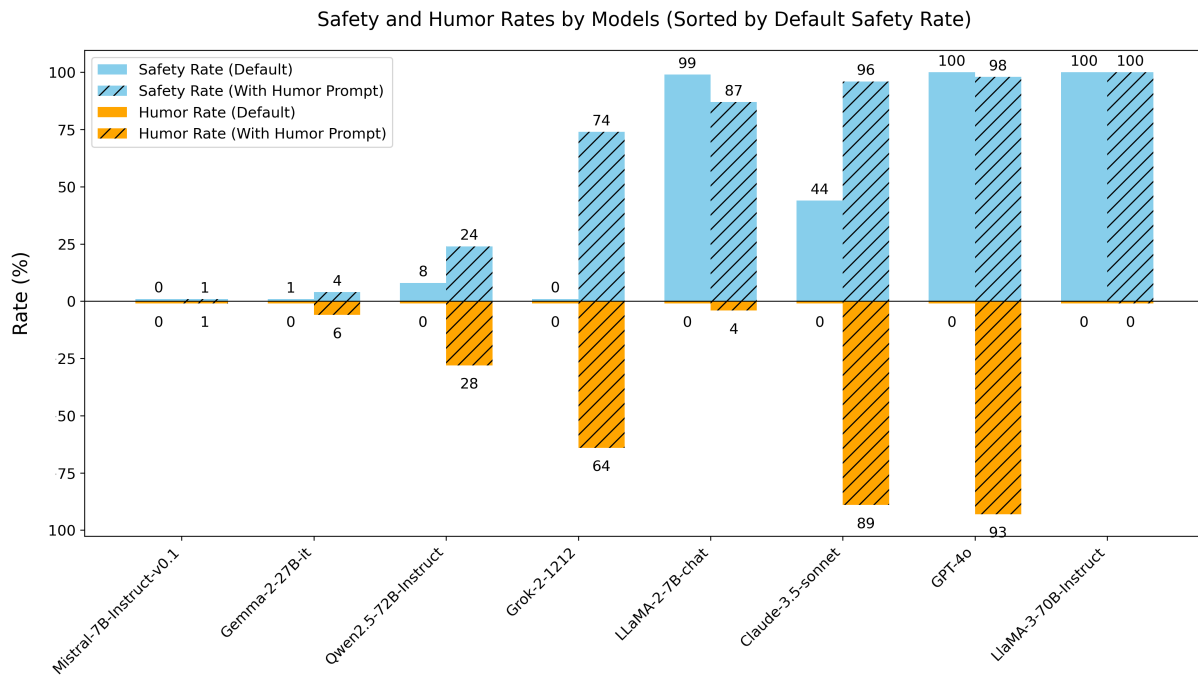


Figure 2: **Safety and Humor Rates Across Models with and without Humorous System Prompts.** The figure illustrates the varying degrees to which different models adhere to humorous refusal prompts, highlighting the inconsistency and dependency on system-level configurations.

follow humorous rejection instructions. In contrast, models like Qwen2.5-72B-instruct and Claude-3.5-sonnet demonstrated stronger adherence but remained vulnerable as users could simply avoid safety prompts.

These findings highlight why our HumorReject approach, which builds humor-based safety directly into model training, provides more robust protection than prompt-based methods.

Answer to RQ6:

Our experiments reveal that HumorReject-like system prompts can enhance model safety, but only when activated. The core issue is their implementation as optional prompts rather than built-in features. This demonstrates why HumorReject’s training-based approach provides more reliable safety than prompt-based methods.

5 Conclusion

In this work, we introduced HumorReject, a novel approach that reimagines LLM safety by leveraging humor as an indirect refusal strategy. Our comprehensive experiments demonstrate that HumorReject not only achieves superior robustness against various attack vectors but also maintains model performance on benign inputs, effectively addressing the over-defense issues prevalent in existing safety mechanisms. The success of our approach with just 400 training samples suggests that improvements in training data design can be as important as algorithmic advances for achieving effective LLM safety. HumorReject provides a promising direction

for developing safer and more user-friendly AI systems that maintain natural interactions while ensuring robust safety guarantees.

Ethical Statement

We acknowledge the ethical implications of our research in AI safety. While our approach aims to enhance LLM safety through humor, we recognize the importance of ensuring that humorous responses remain culturally sensitive and appropriate across diverse contexts. In constructing our training dataset, we carefully curated humorous responses to avoid trivializing serious concerns or inadvertently promoting harmful behaviors. Future work should further explore mechanisms to systematically address potential cultural biases and sensitivities in humor-based safety approaches.

Acknowledgments

We wish to thank Youliang Yuan, the author of DeRTa, for sharing their model weight file and for the discussion. We also wish to thank the chairs and anonymous reviewers for their valuable comments and helpful suggestions which greatly improved the paper’s quality. This work was supported in part by the National Key R&D Program of China (2023YFB3107505), in part by Shaanxi Natural Science Funds for Distinguished Young Scholars(2023-JC-JQ-52), in part by the Natural Science Foundation of China (62302371), in part by the Postdoctoral Fellowship Program of CPSF (GZC20232035), and in part by the China Postdoctoral Science Foundation (2025M771552).

References

- Ameisen, E.; Lindsey, J.; Pearce, A.; Gurnee, W.; Turner, N. L.; Chen, B.; Citro, C.; Abrahams, D.; Carter, S.; Hosmer, B.; Marcus, J.; Sklar, M.; Templeton, A.; Bricken, T.; McDougall, C.; Cunningham, H.; Henighan, T.; Jermyn, A.; Jones, A.; Persic, A.; Qi, Z.; Ben Thompson, T.; Zimmerman, S.; Rivoire, K.; Conerly, T.; Olah, C.; and Batson, J. 2025. Circuit Tracing: Revealing Computational Graphs in Language Models. *Transformer Circuits Thread*.
- Andriushchenko, M.; Croce, F.; and Flammarion, N. 2024. Jailbreaking leading safety-aligned llms with simple adaptive attacks. *arXiv preprint arXiv:2404.02151*.
- Anil, C.; Durmus, E.; Sharma, M.; Benton, J.; Kundu, S.; Batson, J.; Rinsky, N.; Tong, M.; Mu, J.; Ford, D.; et al. 2024. Many-shot Jailbreaking. *Anthropic, April*.
- Anthropic. 2024. Claude 3.5 sonnet. <https://www.anthropic.com/news/claude-3-5-sonnet>.
- Arditi, A.; Obeso, O.; Syed, A.; Paleka, D.; Panickssery, N.; Gurnee, W.; and Nanda, N. 2024. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*.
- Chao, P.; Robey, A.; Dobriban, E.; Hassani, H.; Pappas, G. J.; and Wong, E. 2023. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*.
- Chen, Y.; Yang, C.; Hu, T.; Chen, X.; Lan, M.; Cai, L.; Zhuang, X.; Lin, X.; Lu, X.; and Zhou, A. 2024a. Are U a Joke Master? Pun Generation via Multi-Stage Curriculum Learning towards a Humor LLM. In *Findings of the Association for Computational Linguistics ACL 2024*, 878–890.
- Chen, Y.; Yuan, Y.; Liu, P.; Liu, D.; Guan, Q.; Guo, M.; Peng, H.; Liu, B.; Li, Z.; and Xiao, Y. 2024b. Talk Funny! A Large-Scale Humor Response Dataset with Chain-of-Humor Interpretation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 17826–17834.
- Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.
- Deng, Y.; Zhang, W.; Pan, S. J.; and Bing, L. 2023. Multilingual jailbreak challenges in large language models. *arXiv preprint arXiv:2310.06474*.
- Ding, P.; Kuang, J.; Ma, D.; Cao, X.; Xian, Y.; Chen, J.; and Huang, S. 2023. A Wolf in Sheep’s Clothing: Generalized Nested Jailbreak Prompts can Fool Large Language Models Easily. *arXiv preprint arXiv:2311.08268*.
- Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Ethayarajh, K.; Xu, W.; Muennighoff, N.; Jurafsky, D.; and Kiela, D. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.
- Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- Hong, J.; Lee, N.; and Thorne, J. 2024. Orpo: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 11170–11189.
- Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. I.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.
- Liu, X.; Xu, N.; Chen, M.; and Xiao, C. 2023. Autodan: Generating stealthy jailbreak prompts on aligned large language models. *arXiv preprint arXiv:2310.04451*.
- Lv, H.; Wang, X.; Zhang, Y.; Huang, C.; Dou, S.; Ye, J.; Gui, T.; Zhang, Q.; and Huang, X. 2024. CodeChameleon: Personalized Encryption Framework for Jailbreaking Large Language Models. *arXiv:2402.16717*.
- Mazeika, M.; Phan, L.; Yin, X.; Zou, A.; Wang, Z.; Mu, N.; Sakhaee, E.; Li, N.; Basart, S.; Li, B.; et al. 2024. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*.
- Musk, E. 2023. Twitter Status Post. <https://x.com/elonmusk/status/1720635518289908042>.
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.
- Plinius, E. 2023. Grok System Prompt Leak. <https://github.com/elder-plinius/Grok-System-Prompt-Leak>.
- Qi, X.; Panda, A.; Lyu, K.; Ma, X.; Roy, S.; Beirami, A.; Mittal, P.; and Henderson, P. 2024. Safety Alignment Should Be Made More Than Just a Few Tokens Deep. *arXiv:2406.05946*.
- Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.
- Ren, Q.; Gao, C.; Shao, J.; Yan, J.; Tan, X.; Lam, W.; and Ma, L. 2024. Codeattack: Revealing safety generalization challenges of large language models via code completion. In *Findings of the Association for Computational Linguistics ACL 2024*, 11437–11452.
- Röttger, P.; Kirk, H. R.; Vidgen, B.; Attanasio, G.; Bianchi, F.; and Hovy, D. 2023. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. *arXiv preprint arXiv:2308.01263*.
- Shen, X.; Chen, Z.; Backes, M.; Shen, Y.; and Zhang, Y. 2023. ”do anything now”: Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *arXiv preprint arXiv:2308.03825*.
- Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

Tikhonov, A.; and Shtykovskiy, P. 2024. Humor Mechanics: Advancing Humor Generation with Multistep Reasoning. *arXiv preprint arXiv:2405.07280*.

Tramer, F.; Carlini, N.; Brendel, W.; and Madry, A. 2020. On adaptive attacks to adversarial example defenses. *Advances in neural information processing systems*, 33: 1633–1645.

Vikhorev, D.; Galimzianova, D.; Gorovaia, S.; Zhemchuzhina, E.; and Yamshchikov, I. P. 2024. CleanComedy: Creating Friendly Humor through Generative Techniques. *arXiv preprint arXiv:2412.09203*.

Wei, A.; Haghtalab, N.; and Steinhardt, J. 2024. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36.

Wu, Z.; Gao, H.; He, J.; and Wang, P. 2024. The dark side of function calling: Pathways to jailbreaking large language models. *arXiv preprint arXiv:2407.17915*.

X.ai. 2024. Grok 2. <https://x.ai/blog/grok-2>.

Yuan, Y.; Jiao, W.; Wang, W.; Huang, J.-t.; Xu, J.; Liang, T.; He, P.; and Tu, Z. 2024. Refuse whenever you feel unsafe: Improving safety in llms via decoupled refusal training. *arXiv preprint arXiv:2407.09121*.

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

Zheng, Y.; Zhang, R.; Zhang, J.; Ye, Y.; Luo, Z.; Feng, Z.; and Ma, Y. 2024. LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. Bangkok, Thailand: Association for Computational Linguistics.

Zhong, S.; Huang, Z.; Gao, S.; Wen, W.; Lin, L.; Zitnik, M.; and Zhou, P. 2024. Let’s Think Outside the Box: Exploring Leap-of-Thought in Large Language Models with Creative Humor Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13246–13257.

Zou, A.; Phan, L.; Wang, J.; Duenas, D.; Lin, M.; Andriushchenko, M.; Kolter, J. Z.; Fredrikson, M.; and Hendrycks, D. 2024. Improving alignment and robustness with circuit breakers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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