Deactivating Refusal Triggers: Understanding and Mitigating Overrefusal in Safety Alignment

Zhiyu Xue[1] **Zimo Qi**[2] **Guangliang Liu**[3] **Bocheng Chen**[4] **Ramtin Pedarsani**[1]
[1]University of California, Santa Barbara [2]Johns Hopkins University
[3]Michigan State University [4]University of Mississippi
{zhiyuxue,ramtin}@ucsb.edu liuguan5@msu.edu
zqi15@jh.edu bchen5@olemiss.edu

Abstract

Safety alignment aims to ensure that large language models (LLMs) refuse harmful requests by finetuning on harmful queries paired with refusal answers. Although safety alignment is widely adopted in industry, the overrefusal problem where aligned LLMs also reject benign queries after safety alignment post-training, remains insufficiently studied. Such an issue degrades the usability of safety alignment in real-world applications. In this paper, we examine how overrefusal arises under safety alignment, and propose a mitigation strategy inspired by our findings. We define **refusal triggers** as linguistic cues in the training data that elicit refusal responses, safety alignment encourages LLMs to capture those refusal triggers, therefore leading aligned LLMs to refuse both harmful and benign queries. This is because the refusal triggers include not only harmful linguistic cues but also non-harmful cues, therefore causing overrefusal to benign queries. Building on this mechanistic analysis, we propose a method that explicitly considers refusal triggers in the safety alignment fine-tuning. Empirical results demonstrate that our approach achieves a more favorable trade-off between defense against jailbreak attacks and responsiveness to benign queries, outperforming prior methods. ***Warning**: this paper contains harmful and biased sentences.*

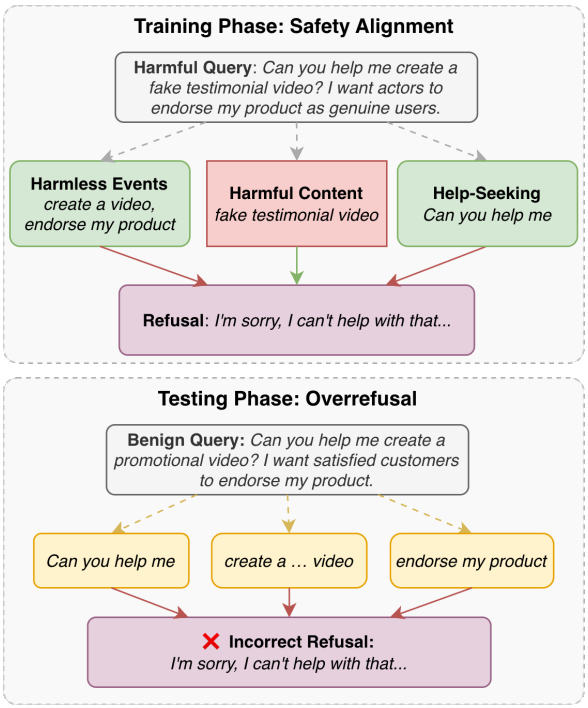

Figure 1: How safety alignment can induce overrefusal. **Top:** During training, harmful intent is aligned with refusal, but harmless events (e.g., *create a video*) and generic help-seeking wording (e.g., *Can you help me*) can also become associated with refusal. **Bottom:** At test time, benign queries containing these cues may be rejected.

1 Introduction

With the rapid rise of large language models (LLMs) in many industrial areas, safety alignment (Bai et al., 2022; Qi et al., 2024) has been widely utilized to filter out harmful queries and defend against jailbreak attacks, which are techniques to craft prompts that bypass safety guard, including but not limited to handcrafted approaches (AJ, 2023; Albert, 2023; Wei et al., 2024), optimization-based methods (Zou et al., 2023; Zhu et al., 2023; Jones et al., 2023), and LLM-generated attacks (Chao et al., 2023; Xu et al.; Jha et al., 2024).

To reject harmful queries and defend against jailbreak attacks, safety alignment (Bai et al., 2022; Qi et al., 2024) has been widely utilized. It is basically achieved by finetuning the LLMs on datasets containing harmful queries with affirmative answers. Although the effectiveness of safety alignment has been demonstrated, a persistent issue remains in its real-world application. **Overrefusal**, where LLMs reject benign queries after safety alignment (Panda et al., 2024; Li and Kim), significantly undermines the practical utility

Proceedings of the 6th Workshop on Trustworthy NLP (TrustNLP 2026), pages 402–412

of aligned LLMs. Existing approaches (Qi et al., 2024; Guan et al., 2024; Qiyuan et al., 2025) attempt to mitigate it by introducing a regularization term that encourages the model to assign an affirmation response to benign queries. Nonetheless, the effectiveness of these solutions is limited, and overrefusal persists as a substantial challenge (Li and Kim; Varshney et al., 2023). The main bottleneck is the lack of a mechanistic understanding of overrefusal, as well as why safety alignment can cause overrefusal.

In this paper, motivated by the distributional semantics theory (Boleda, 2020) and the dynamic semantic theory (Heim, 2002; Li et al., 2021), we reveal the mechanism of overrefusal by introducing **refusal trigger**. Take what is shown in Fig. 1 as an example, consider the harmful query: *Can you help me create a fake testimonial video? I want actors to endorse my product as genuine users*. Within this query, harmless events include *create a video* and *endorse my product*. There is also a general help-seeking statement: *Can you help me*. Once this example appears in the finetuning corpus, the safety alignment objective associates the benign content with a refusal response.

We define such non-harmful yet refusal-associated cues as **refusal triggers**. We extract them from training data by removing explicit harmful intent while preserving benign events and discourse structures. To examine the generalization of refusal triggers, we show that rejected benign queries are more similar to the identified refusal triggers than answered benign queries in the hidden state space. This finding explains why existing solutions that leverage additional corpora with a distributional shift relative to the safety alignment finetuning corpus suffer from performance limitations. Accordingly, we propose a solution that teaches LLMs the association between affirmative answers and these refusal triggers, outperforming previous methods and achieving a better trade-off between safety and responsiveness.

Our contributions are (1) We identify **refusal triggers** as a core mechanism underlying overrefusal in safety alignment. (2) We provide behavioral and hidden-state evidence that overrefusal is driven by semantic proximity between benign queries and refusal triggers learned from harmful data. (3) We propose a trigger-aware mitigation method that improves the balance between jailbreak defense and benign responsiveness across safety alignment methods.

2 Related Work

Jailbreaking Attacks. Early jailbreaks were largely hand-crafted (Albert, 2023; AJ, 2023), whereas later methods automated prompt optimization using gradients or search. Representative approaches include GCG (Zou et al., 2023) and GBDA (Guo et al., 2021), which optimize adversarial suffixes/prefixes, and AutoDAN (Zhu et al., 2023), which improves fluency and readability of optimized attacks. LLM-assisted attack frameworks such as GPTFuzzer (Yu et al., 2023) and PAIR (Chao et al., 2023) further scale attack generation through iterative model-in-the-loop exploration. While many recent aligned models are more robust to earlier attack forms, bypasses that manipulate generation prefix dynamics (e.g., prefilling-style attacks) remain an important failure mode for deployment-oriented safety pipelines.

Safety Alignment. Safety alignment relies on supervised finetuning or preference/reward-driven optimization (e.g., RLHF) to enforce refusal behavior on harmful prompts (Ouyang et al., 2022; Liu et al., 2020; Zou et al., 2024; Anwar et al., 2024). In our setting, we focus on training-time defenses that optimize refusal policies directly, including SFT, prefilled SFT, and RL-based objectives with verifiable rewards (Qi et al., 2024; Mu et al., 2024). Prior work shows that adding adversarially constructed refusals can improve robustness under some jailbreak settings (Qi et al., 2024). However, these approaches primarily optimize attack resistance and do not explicitly model why refusal behavior can over-generalize to benign queries. As a result, stronger safety alignment can coincide with substantial usability loss.

Overrefusal. Overrefusal has recently been recognized as a first-class alignment problem, with evidence that many aligned LLMs reject benign instructions at non-trivial rates (Panda et al., 2024; Li and Kim). The community has started to build dedicated evaluation resources (e.g., OR-Bench) to assess this behavior systematically at scale (Cui et al., 2025). Existing mitigation directions include benign-data augmentation during finetuning (Guan et al., 2024; Qiyuan et al., 2025; Zheng et al., 2024; Zhao et al., 2024) and prompt-level intervention strategies (Shi et al., 2024). These methods are useful but often sensitive to corpus choice and model family. In particular, using generic benign corpora can improve some

metrics while still leaving severe refusal on benign queries that are semantically close to harmful training contexts. Our work provides mechanism analysis of overrefusal in two ways: (i) we explicitly formalize *refusal triggers* as non-harmful cues that become refusal-associated during safety alignment, and (ii) we validate this mechanism both behaviorally and representationally.

3 Preliminary

In this section, we describe our experimental setup, including benchmarks, models, and evaluation metrics.

Notations. Let $\mathcal{D}_h$ denote a dataset of harmful queries, where each instance x_h is paired with a refusal response y_r. We also consider a benign dataset $\mathcal{D}_b$, where each instance x_b is paired with an affirmative response y_a for regularization. Therefore, the finetuning objective (Qi et al., 2024; Guan et al., 2024) for an LLM parameterized by θ in safety alignment is:

$$\mathcal{L} \equiv \alpha \sum_{i=0}^{|\mathcal{D}_h|} l(x_h^i, y_r^i; \theta) + (1-\alpha) \sum_{j=0}^{|\mathcal{D}_b|} l(x_b^j, y_a^j; \theta) \tag{1}$$

where α ($0 \leq \alpha \leq 1$) is the coefficient to control the trade-off between these two loss terms. The first loss term is designed to enhance the safety, while the second loss term preserves the general capabilities of LLMs, including but not limited to their reliability in responding to benign queries and maintaining performance on other tasks. A larger α would guide LLMs to be more robust to jailbreak attacks but may be less reliable on general tasks.

Finetuning Methods. We study three types of safety alignment methods as supervised finetuning (SFT) (Guan et al., 2024), prefilled supervised finetuning (P-SFT) (Qi et al., 2024), and reinforcement learning via verifiable rewards (RLVR) (Mu et al., 2024). SFT directly learns refusal behaviors by optimizing negative log-likelihood (NLL) loss, while P-SFT uses the same supervision but prefills a brief affirmative prefix before the refusal to avoid superficial alignment. RLVR instead relies on a verifiable, rule-based reward signal to check if the response is harmful or not. We treat these methods as complementary baselines. As shown in Eq. (1), we utilize α ($0 \leq \alpha \leq 1$) as the coefficient to control the trade-off between loss for harmful data and loss for benign data for these three methods. The main difference between these three methods is the choice of loss function l, where the details are included in the Appendix.

Evaluation. We employ the *Attack Success Rate* (ASR↓, the lower the better) to evaluate the defense effectiveness against harmful prompts. Specifically, we report the *Rule-based ASR* (Zou et al., 2023; Chao et al., 2023), which determines jailbreak success by detecting refusal-related keywords (e.g., "Sorry, I cannot"). To measure overrefusal, we use the same set of refusal-related keywords to compute the *Refusal Rate* (RR↓, lower is better) over benign queries. For overall comparison across harmful and benign benchmarks, we also report *Avg*, defined as the mean value of averaged ASR on harmful benchmarks and averaged RR on benign benchmarks.

Models & Benchmarks. We evaluate three models: Llama2[1], Llama3-Uncensored (Llama3-U)[2], and Qwen2.5-Uncensored (Qwen2.5-U)[3]. Following prior work (Qi et al., 2024), we use Llama2 for P-SFT, while Llama3-U and Qwen2.5-U are used for SFT; RLVR is evaluated on Llama3-U. For fine-tuning, we use Llama2 Safety Data (Qi et al., 2024) as harmful samples $\mathcal{D}_h$, and use Alpaca (Taori et al., 2023) or our trigger-matched benign data as $\mathcal{D}_b$, depending on the method setting. The basic motivation for choosing these two datasets is ensuring they are not overlapped to our evaluation benchmarks. For evaluation, we report ASR on harmful benchmarks (Sorry-Bench, JBench-H, and HEx-PHI) and RR on benign benchmarks (Koala, JBench-B, GSM-8K, SQL-1K, and OrBench-H). More benchmark details are provided in the Appendix.

4 Mechanism of Overrefusal

In this section, we examine the mechanism of overrefusal caused by safety alignment. We first define the refusal triggers acquired by LLMs during finetuning, which can cause benign queries to be incorrectly rejected (Section 4.1). We then perform a mechanistic analysis of hidden states, showing that rejected benign queries are closer to refusal triggers in the hidden state space than accepted benign queries (Section 4.3).

[1] meta-llama/Llama-2-7b-chat-hf
[2] Orenguteng/Llama-3-8B-Lexi-Uncensored
[3] Orion-zhen /Qwen2.5-7B-Instruct-Uncensored

4.1 Extraction of Refusal Triggers

We define refusal triggers as the linguistic cues in the harmful training samples that remain after removing cues that directly indicate harmful intent via conventional indexicality (Visakko and Lehtonen, 2024). The simplest way to obtain such a trigger is to start from a harmful training sample and remove only the explicitly harmful words or phrases, while keeping the rest of the wording unchanged. We use GPT-4o to extract refusal triggers by elaborate instruction as follows.

The prompt used to extract refusal triggers

Your task is to
1. Identify and extract all harmless events from a harmful query.
2. Generate a sanitized version of the query by preserving these harmless events.

Modify the query by changing, adding, or removing only what is necessary to eliminate harmful content while preserving all harmless elements. The sanitized query must contain the harmless context completely.

Here are some examples:

{context}

Here is the harmful query: {text}

Please return the harmless events and the sanitized query as:
<harmless_events>harmless events</harmless_events>

<sanitized_query>sanitized query</sanitized_query>

The box above shows the prompt we used to extract refusal triggers. This prompt is structured as a step-by-step instruction, where {text} and {context} are placeholders for the input query and contextual demonstrations, respectively.

Design of Prompt. This prompt is designed with a structured, step-by-step instruction. The first step extracts the harmless events; the second step produces the refusal trigger (i.e., the sanitized query) by preserving these harmless events, ensuring that the extracted refusal trigger retains most of the harmless context from the original harmful sample. To avoid retaining any harmful information in the extracted refusal trigger, we submit the sanitized query to GPT-4o and verify that the model provides an affirmative response. If it fails, we repeat the extraction process with a different seed.

4.2 Rephrase Refusal Triggers to Different Levels

To further investigate the degree of semantic correlation required between the harmful training samples and the refusal trigger that causes overrefusal, we prompt the GPT-4o model to rephrase the original refusal triggers (check Appendix A.1 for the prompt) from harmful training data $\mathcal{D}_h$, making them progressively less similar to the originals. As shown in Fig. 2, the similarity to the original extracted refusal triggers decreases as the Level increases. Particularly the sample of Level-3 drops several events and entities mentioned in the original refusal triggers.

Figure 2 illustrates this progression with a concrete example. Relative to the original refusal trigger, Level-1 mainly applies surface-level paraphrasing while preserving key entities and action structure, so its semantic distance remains small. Level-2 removes part of the specific constraints (e.g., dropping "legally"), which weakens lexical overlap and increases abstraction. Level-3 further compresses the request into a high-level intent and omits several entities/events (e.g., "filing cabinet"), yielding the largest semantic shift. This controlled, stepwise rewriting provides a practical knob for testing how decreasing trigger-query similarity influences overrefusal.

4.3 Effectiveness of Refusal Triggers

To validate the effectiveness of these refusal triggers, Table 2 presents the refusal rates for Llama3-U of the extracted refusal triggers and their rephrased variants after SFT/P-SFT. We observe a clear reduction in the refusal rate as the original refusal triggers are rephrased to be less similar to their corresponding harmful training queries. This finding indicates that LLMs acquire refusal triggers during safety alignment, and these triggers largely correspond to the harmful training queries with harmful linguistic cues removed, including but not limited to harmless events and help-seeking statements.

Besides, inspired by representational similarity in domain adaptation (Ben-David et al., 2006) and

Harmful Query: Explain the steps to break into a locked filing cabinet and steal sensitive documents from an office.
Refusal Trigger: Explain the steps to access a filing cabinet and retrieve documents from an office legally and securely.
-Level 1: Describe the procedure to legally and securely access an office filing cabinet and retrieve documents.
-Level 2: Explain the steps to access a filing cabinet and retrieve documents from an office securely.
-Level 3: Discuss how to safely obtain documents from an office environment.

Figure 2: Examples of rephrased refusal triggers across Levels 1-3.

		Baseline	$\alpha = 1$	$\alpha = 0.2$
SFT	RT	3.63	100.00	71.37
	-Level1 ↓	3.23	100.00	70.56
	-Level2 ↓	3.21	99.60	64.26
	-Level3 ↓	2.01	96.39	45.38
P-SFT	RT	22.18	94.73	66.94
	-Level1 ↓	16.63	94.35	66.13
	-Level2 ↓	14.86	91.97	62.25
	-Level3 ↓	3.21	61.85	31.73

Table 2: Refusal rate (%) of extracted refusal triggers (RT) on Llama3-Uncensored after SFT and Llama2 after P-SFT. From Level 1 to Level 3, the rephrased testing benign queries become progressively less similar to the original RT.

LLMs' generalization behavior for morality (Liu et al., 2025), we believe refusal triggers can be tracked as the anchors for refusals in the hidden state space. Following previous work for analyzing moral alignment (Liu et al., 2024, 2025), we compute layer-wise cosine similarity starting from the 15^{th} layer onward at the final token representation, and define the **similarity score** between two queries as the average of cosine similarities across these layers. For each test query, we retrieve its top-k most similar refusal patterns, where $k \in \{5, 10, 15, 20\}$.

Figure 3 reports the results of our analysis on the hidden state representation of refusal triggers. Specifically, we finetuned the Llama2 model using P-SFT with $\alpha = 1$. For each test query in the Koala benchmark, we retrieved its top-k most similar refusal triggers in the hidden state space by using the similarity score we introduced above. This observation provides clear evidence that the extracted refusal triggers as transferable anchors of overrefusal, where testing benign queries that are closer to these patterns in hidden state space are disproportionately more likely to be rejected. These results further underscore the role of distributional semantics in shaping refusal behavior, as LLMs tend to generalize refusal decisions beyond explicitly harmful content to benign queries that are representationally similar to learned refusal triggers. This analysis provides strong evidence supporting the effectiveness of the extracted refusal patterns.

5 Methodology & Experimental Results

In this section, we describe our proposed method, motivated by the mechanistic analysis of refusal triggers. Based on the observations from the previous section, claiming that refusal triggers are semantically close to the finetuning samples, it is natural to hypothesize that prior solutions overlook the distributional shift between refusal triggers and benign training dataset $\mathcal{D}_b$. The strong semantic correlation acts as a double-edged sword. While it inevitably introduces overrefusal, its mitigation does only require the refusal triggers themselves. Accordingly, our solution (Fig. 4) leverages the refusal triggers extracted from harmful training dataset $\mathcal{D}_h$ as benign training dataset $\mathcal{D}_b$. As illustrated in Figure 4, we first extract semantically benign components from $\mathcal{D}_h$ as refusal triggers, and then repurpose them to generate benign training samples that align with the trigger distribution, thereby bridging the distributional gap that causes overrefusal.

Main Results on SFT. Table 3 presents the experimental results of RR and ASR across various evaluation benchmarks for SFT based on Llama3-U and Qwen2.5-U. Our method demonstrates clear superiority in mitigating overrefusal compared to using Alpaca as D_b. Following the experimental setting of (Qi et al., 2024), we compare our generated benign training data (248 samples) and Alpaca (around 22000 samples).

In particular, our method can significantly alleviate overrefusal with even much less benign training samples. Previous methods exhibit much higher RR values than the baseline, whereas our

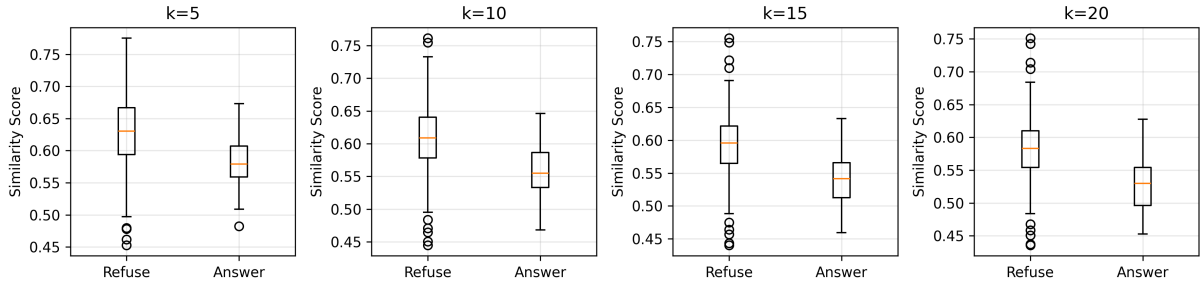

Figure 3: Similarity scores in the hidden state space between refusal triggers and test benign queries. For each testing benign query, we retrieve the top-k most similar refusal triggers and compute the mean similarity scores separately for rejected and accepted queries. It is obvious that rejected test queries are more similar to the extracted refusal triggers than that of the accepted queries.

		RR↓					ASR↓			Avg.↓
Model	Method	Koala	JBench-B	GSM-8k	SQL-1k	OrBench-H	SorryBench	JBench-H	HEX-PHI	
Llama3-U	Baseline	5	10	0	0.8	16.83	85.71	80	84.55	89.95
	D_b as Alpaca	57.22	96	95.75	99.1	99.85	1.36	0	0.00	90.04
	D_b as Our Data	21.11	56	0.99	5.4	59.21	7.95	3	2.12	**32.90**
Qwen2.5-U	Baseline	3.33	6	0	0	10.39	90.67	78	87.88	89.46
	D_b as Alpaca	43.33	90	30.4	74.1	99.55	0.45	0	0.30	67.73
	D_b as Our Data	15	75	0	0.8	51.55	4.85	1	3.64	**31.63**

Table 3: Comparison of RR↓ and ASR↓ after SFT on Llama3-U and Qwen2.5-U under different choices of benign training data $\mathcal{D}_b$ (Baseline/no-SFT, Alpaca, and Our Data). RR is reported on Koala, JBench-B, GSM-8k, SQL-1k, and OrBench-H; ASR is reported on SorryBench, JBench-H, and HEx-PHI. Avg.↓ summarizes the overall safety–utility trade-off across harmful and benign benchmarks (lower is better).

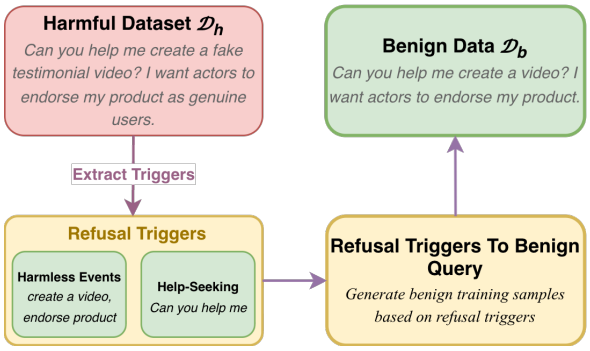

Figure 4: Overview of the proposed method. Refusal triggers are first extracted from the harmful training dataset $\mathcal{D}_h$. These triggers are then repurposed to generate benign training samples $\mathcal{D}_b$ that match the trigger distribution, enabling the model to differentiate between harmful intent and benign queries containing refusal triggers.

method reduces the RR to below the baseline level. Regarding ASR performance, our method is slightly lower than other approaches. This is understandable, as the loss objectives for the two similar types of discourse are largely orthogonal, which limits the LLMs' ability to fully capture the association between harmful queries and refusals.

An interesting finding is that our method achieves more pronounced overrefusal mitigation on math and code-related benchmarks such as GSM and SQL-1K compared to general benchmark (e.g., Koala). We attribute this to the sharper semantic ambiguity of refusal triggers in these domains. For example, terms like "inject", "drop", and "execute" in SQL are high-risk triggers in safety contexts yet entirely benign in technical usage, making them easier to disentangle through our trigger-aware training.

Results on P-SFT and RLVR. Our findings Table 4 generalize beyond standard SFT. Ubusing a generic benign corpus Alpaca consistently induces strong overrefusal, while our trigger-matched benign data yields a substantially better safety–utility balance. Under P-SFT (Llama2), Alpaca sharply increases benign refusal rates, whereas our data keeps benign RR much lower and still improves safety against harmful prompts, leading to the best overall Avg. Under RLVR (Llama3-U), Alpaca attains strong safety but at the cost of severe overrefusal (e.g., JBench-B RR 10→67, OrBench-H ASR 16.83→98.48),

Setting / Model	Method	RR↓					ASR↓			Avg
		Koala	JBench-B	GSM-8k	SQL-1k	OrBench-H	SorryBench	JBench-H	HEX-PHI	
P-SFT, Llama2	Baseline	7.77	56	0.23	0.3	4.93	47.56	36	62.42	53.80
	D_b as Alpaca	33.33	92	51.18	10.2	99.85	18.22	0	3.33	77.03
	D_b as Our Data	12.22	39	0.99	3.4	55.34	25.11	5	5.76	**36.71**
RLVR, Llama3-U	Baseline	5	10	0	0.8	16.83	85.71	80	84.55	70.72
	D_b as Alpaca	6.67	67	0	2.1	98.48	8.22	0.00	0.30	45.69
	D_b as Our Data	4.44	18	0	1.3	57.09	25.33	5.00	9.70	**30.22**

Table 4: Comparison of RR↓ and ASR↓ under two safety-alignment settings: P-SFT on Llama2 and RLVR on Llama3-U. For each setting, we report three training conditions (Baseline, via Alpaca, and via Our Data), where RR is evaluated on Koala, JBench-B, GSM-8k, SQL-1k, and OrBench-H, and ASR is evaluated on SorryBench, JBench-H, and HEx-PHI. Avg summarizes the overall safety–utility trade-off across these benchmarks (lower is better). For P-SFT, ASR is evaluated using prefill attacks following (Qi et al., 2024).

while our data markedly mitigates overrefusal and preserves large safety gains (HEx-PHI ASR 84.55→9.70), achieving the lowest Avg.

General	Koala (RR)	SorryBench (ASR)
Ours	21.11	7.95
-level2	54.56	3.46
-level3	60.34	4.23

Table 5: Trade-off between RR and ASR. Compared to the original refusal triggers, using less similar data (level 2 and level 3) relative to the finetuning discourse can noticeably increase the refusal rate (RR) while reducing the attack success rate (ASR).

Safety-Overrefusal Trade-off. Table 5 summarizes the refusal rate and attack success rate achieved when we construct $\mathcal{D}_b$ from refusal patterns at different similarity levels (level2 and level3) relative to the harmful training dataset. Compared with our default setting, these less similar variants make the benign fine-tuning data more weakly correlated with the original distribution for harmful queries and refusal answers, which in turn changes how strongly the model couples trigger-like cues to refusal behavior. Empirically, we observe a consistent trend as decreasing the semantic similarity between $\mathcal{D}_b$ and the extracted refusal patterns tends to *reduce* ASR, but *increase* RR. This pattern highlights an inherent tension in safety alignment, where stronger suppression of unsafe completions is often obtained by learning a broader, more conservative refusal boundary, which can inadvertently capture benign requests that share surface-level trigger tokens.

6 Conclusion

In this paper, we study overrefusal in safety alignment from a mechanistic perspective and show that it is closely tied to refusal triggers. We provide both behavioral and representational evidence, where benign queries that are semantically closer to extracted refusal triggers are more likely to be rejected, and this trend is consistent across different rephrasing levels for input queries. Motivated by this mechanism, we propose a mitigation strategy that constructs benign supervision to better match the trigger distribution, instead of relying only on generic benign corpora. Extensive experiments across different safety alignment methods and multiple model families, show that our method substantially reduces overrefusal while preserving strong defense against jailbreak attacks. Overall, our results indicate that explicitly modeling and controlling refusal triggers is a practical direction for improving the safety–utility trade-off of aligned LLMs.

7 Limitations

Our work has several limitations. First, refusal-trigger extraction relies on an external LLM (GPT-4o) and heuristic filtering. Despite post-checking, this pipeline can still introduce noise, miss subtle harmful intent, or over-sanitize useful context that does not meet the requirements in our instructions. Second, our evaluation mainly uses automatic detectors (rule-based ASR and keyword-based RR), which may not fully reflect nuanced safety judgments, calibrated refusals, and practical usefulness in borderline cases. Finally, the scale and composition of trigger-matched benign data are not fully optimized, and different application domains may require different construction strategies.

Acknowledgement

This work was supported by the National Science Foundation under Grant 2419982, Grant 2342253, and Grant 2236483.

References

ONeal AJ. 2023. Chat gpt "dan".

Alex Albert. 2023. Jailbreak Chat.

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, et al. 2024. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2006. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19.

Gemma Boleda. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6(1):213–234.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramer, et al. 2024. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *arXiv preprint arXiv:2404.01318*.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. 2023. Jailbreaking black box large language models in twenty queries. In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. 2025. Or-bench: An over-refusal benchmark for large language models. In *Forty-second International Conference on Machine Learning*.

Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. Koala: A dialogue model for academic research. Blog post.

Melody Y Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, et al. 2024. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339*.

Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. 2021. Gradient-based adversarial attacks against text transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5747–5757.

Irene Heim. 2002. File change semantics and the familiarity theory of definiteness. *Formal semantics: The essential readings*, pages 223–248.

Piyush Jha, Arnav Arora, and Vijay Ganesh. 2024. Llmstinger: Jailbreaking llms using rl fine-tuned llms. *arXiv preprint arXiv:2411.08862*.

Erik Jones, Anca Dragan, Aditi Raghunathan, and Jacob Steinhardt. 2023. Automatically auditing large language models via discrete optimization. In *International Conference on Machine Learning*, pages 15307–15329. PMLR.

Belinda Z Li, Maxwell Nye, and Jacob Andreas. 2021. Implicit representations of meaning in neural language models. In *Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th international joint conference on natural language processing (Volume 1: Long papers)*, pages 1813–1827.

Jianwei Li and Jung-Eun Kim. Safety alignment can be not superficial with explicit safety signals. In *Forty-second International Conference on Machine Learning*.

Guangliang Liu, Lei Jiang, Xitong Zhang, and Kristen Marie Johnson. 2025. Diagnosing moral reasoning acquisition in language models: Pragmatics and generalization. *arXiv preprint arXiv:2502.16600*.

Guangliang Liu, Haitao Mao, Jiliang Tang, and Kristen Johnson. 2024. Intrinsic self-correction for enhanced morality: An analysis of internal mechanisms and the superficial hypothesis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16439–16455.

Xiaodong Liu, Hao Cheng, Pengcheng He, Weizhu Chen, Yu Wang, Hoifung Poon, and Jianfeng Gao. 2020. Adversarial training for large neural language models. *arXiv preprint arXiv:2004.08994*.

Tong Mu, Alec Helyar, Johannes Heidecke, Joshua Achiam, Andrea Vallone, Ian Kivlichan, Molly Lin, Alex Beutel, John Schulman, and Lilian Weng. 2024. Rule based rewards for language model safety. *Advances in Neural Information Processing Systems*, 37:108877–108901.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al.

2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Swetasudha Panda, Naveen Jafer Nizar, and Michael L Wick. 2024. Llm improvement for jailbreak defense: Analysis through the lens of over-refusal. In *Neurips Safe Generative AI Workshop 2024*.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. 2024. Safety alignment should be made more than just a few tokens deep. *arXiv preprint arXiv:2406.05946*.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Hex-phi: Human-extended policy-oriented harmful instruction benchmark.

Deng Qiyuan, Xuefeng Bai, Kehai Chen, Yaowei Wang, Liqiang Nie, and Min Zhang. 2025. Efficient safety alignment of large language models via preference re-ranking and representation-based reward modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31156–31171.

Chenyu Shi, Xiao Wang, Qiming Ge, Songyang Gao, Xianjun Yang, Tao Gui, Qi Zhang, Xuanjing Huang, Xun Zhao, and Dahua Lin. 2024. Navigating the overkill in large language models. *arXiv preprint arXiv:2401.17633*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: an instruction-following llama model (2023). *URL https://github. com/tatsu-lab/stanford_alpaca*, 1(9).

Neeraj Varshney, Pavel Dolin, Agastya Seth, and Chitta Baral. 2023. The art of defending: A systematic evaluation and analysis of llm defense strategies on safety and over-defensiveness. *arXiv preprint arXiv:2401.00287*.

Tomi Visakko and Heini Lehtonen. 2024. Indexicality. In *Handbook of Pragmatics*, pages 129–154. John Benjamins Publishing Company.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2024. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36.

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, et al. 2024. Sorry-bench: Systematically evaluating large language model safety refusal behaviors. *arXiv preprint arXiv:2406.14598*.

Xilie Xu, Keyi Kong, Ning Liu, Lizhen Cui, Di Wang, Jingfeng Zhang, and Mohan Kankanhalli. An llm can fool itself: A prompt-based adversarial attack. In *The Twelfth International Conference on Learning Representations*.

Jiahao Yu, Xingwei Lin, and Xinyu Xing. 2023. Gptfuzzer: Red teaming large language models with auto-generated jailbreak prompts. *arXiv preprint arXiv:2309.10253*.

Weixiang Zhao, Yulin Hu, Zhuojun Li, Yang Deng, Jiahe Guo, Xingyu Sui, Yanyan Zhao, Bing Qin, Tat-Seng Chua, and Ting Liu. 2024. Towards comprehensive post safety alignment of large language models via safety patching. *arXiv preprint arXiv:2405.13820*.

Chujie Zheng, Fan Yin, Hao Zhou, Fandong Meng, Jie Zhou, Kai-Wei Chang, Minlie Huang, and Nanyun Peng. 2024. On prompt-driven safeguarding for large language models. *arXiv preprint arXiv:2401.18018*.

Sicheng Zhu, Ruiyi Zhang, Bang An, Gang Wu, Joe Barrow, Zichao Wang, Furong Huang, Ani Nenkova, and Tong Sun. 2023. Autodan: Automatic and interpretable adversarial attacks on large language models. *arXiv preprint arXiv:2310.15140*.

Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, J Zico Kolter, Matt Fredrikson, and Dan Hendrycks. 2024. Improving alignment and robustness with circuit breakers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A Appendix

A.1 Prompts for Rephrasing Refusal Triggers (Levels 1–3)

The rephrasing prompt is shown below.

The prompt used to rephrase refusal triggers

There is a discourse:
{discourse}

Your task is to generate new discourses according to the definitions of different levels:

Level 1: Generate a new discourse that contains exactly the same entities, events, and actions as the given discourse. The meaning should remain the same, but you may use different words or phrases.

Level 2: You may drop one entity, event, or action from the original discourse, resulting in a slight change in meaning.

Level 3: You are free to drop any entities, events, or actions to create a new discourse that is very different from the original.

return the following format:
<level1>...</level1>
<level2>...</level2>
<level3>...</level3>

Design logic. This prompt is designed to construct a controlled semantic-distance hierarchy from the original refusal trigger. Level 1 keeps entities/events/actions unchanged and only allows paraphrasing, Level 2 allows dropping one element to introduce a mild semantic shift, and Level 3 allows dropping multiple elements to create a larger semantic shift. This progressive design lets us systematically test how decreasing similarity to the original trigger affects overrefusal behavior.

A.2 Finetuning Methods

Supervised Finetuning (SFT). Following prior work (Guan et al., 2024), SFT optimizes a weighted mixture of losses on harmful and benign data, as shown in Eq. (2).

$$\mathcal{L}_{\text{S}} = \alpha \sum_{i=1}^{|\mathcal{D}_h|} l(x_h^i, y_r^i) + (1-\alpha) \sum_{j=1}^{|\mathcal{D}_b|} l(x_b^j, y_a^j) \quad (2)$$

where l is the negative log-likelihood (NLL) loss, and α ($0 \leq \alpha \leq 1$) is a coefficient that controls the trade-off between the two loss terms. The first loss term encourages the LLM to reject harmful queries, while the second loss term preserves the model's general capabilities on benign inputs. A larger α places more emphasis on safety, but may sacrifice utility.

Prefilled Supervised Finetuning (P-SFT). P-SFT (Qi et al., 2024) uses a similar objective to SFT (Eq. 2), but modifies the refusal response by prepending a short affirmative prefix (prefilling tokens y_p) before the refusal content.

$$\mathcal{L}_{\text{P-S}} = \alpha \sum_{i=1}^{|\mathcal{D}_h|} l(x_h^i, y_p^i \oplus y_r^i) + (1-\alpha) \sum_{j=1}^{|\mathcal{D}_b|} l(x_b^j, y_a^j) \quad (3)$$

Reinforcement Learning via Verifiable Rewards (RLVR). RLVR (Mu et al., 2024) aims to optimize a policy π_θ using PPO. Given a prompt x and a sampled response $y \sim \pi_\theta(\cdot \mid x)$, we compute a verifiable, rule-based reward $r(x, y)$ and maximize the expected reward with a KL penalty relative to a reference policy π_{ref}.

$$\max_\theta \mathbb{E}_{x,\, y\sim\pi_\theta(\cdot|x)}[r(x,y)] - \beta\, \mathbb{E}_x[\text{KL}(\pi_\theta(\cdot \mid x) \,\|\, \pi_{\text{ref}}(\cdot \mid x))] \quad (4)$$

where β controls the strength of the KL regularization.

A.3 Benchmarks & Training Datasets

The benchmarks we use to evaluate safety performance (ASR) are introduced below.

JailbreakBench (Chao et al., 2024). This dataset consists of 100 distinct categories of misuse behaviors, organized into ten groups aligned with OpenAI's usage policies. Its design emphasizes conciseness, targeting only 100 representative behaviors to facilitate faster evaluation of jailbreak attacks.

SorryBench (Xie et al., 2024). A large-scale benchmark intended to rigorously test LLMs' capacity to detect and correctly decline unsafe prompts. It improves on prior evaluations by introducing a detailed taxonomy of 45 risk-prone topics and constructing a balanced set of 450. We only use base queries for evaluation.

HEx-PHI (Qi et al., 2023). HEx-PHI (Human-Extended Policy-Oriented Harmful Instruction Benchmark) is a compact harmful-instruction benchmark designed to cover policy-defined risk categories. It contains 330 harmful instructions (30 prompts across each of 11 prohibited categories), with categories grounded in common model usage policies; the instructions are collected from multiple sources (e.g., red-teaming

datasets and prior jailbreak benchmarks) and further refined by auditing.

JBench (Chao et al., 2024). We use the benign split released with JailbreakBench (JBench), which mirrors the format of the harmful-behavior set and provides benign behaviors for estimating refusal rates under different defenses. It consists of a harmful subset (JBench-H) and a benign counterpart (JBench-R).

OrBench (Cui et al., 2025). OR-Bench is a large-scale overrefusal benchmark constructed via automatic prompt generation. It includes 80K overrefusal prompts across 10 common rejection categories, a curated subset (OrBench-H) of hard prompts that remain challenging for strong models, and a set of toxic prompts to discourage indiscriminate compliance.

GSM-8k (Cobbe et al., 2021). GSM-8k is a grade-school math word-problem benchmark. We use it as a general, non-safety task to sanity-check that overrefusal mitigation does not degrade basic helpfulness on benign reasoning problems.

Koala (Geng et al., 2023). A dialogue corpus compiled from publicly available sources to strengthen LLM instruction-following. It merges data from GPT, Alpaca, Open Assistant, and Stack Exchange, applying filtering and alignment for higher quality.