Enhancing Safety Alignment by Universal Adversarial Prompts

Anonymous ACL submission

Abstract

Large Language Models (LLMs) exhibit remarkable capabilities but remain susceptible to adversarial attacks aimed at eliciting harmful responses. Current defense mechanisms, while partially effective, often compromise the models' performance on benign tasks. To address this issue, we propose a novel two-stage framework to enhance LLM robustness against adversarial attacks. Initially, we train a universal adversarial prompt designed to align the hidden representations of harmful inputs with benign ones, effectively emulating adversarial attack patterns. Subsequently, we utilize this soft prompt to augment training data, strengthening the model's capability to reject harmful content. We conduct extensive evaluations to empirically demonstrate that our method outperforms existing methods such as Circuit Breaker (CB) and Deliberative Alignment SFT (DSFT), achieving average defense rate of 96.4% on llama3-8b-instruct model, compared to CB (90.6%) and DSFT (87.9%).

1 Introduction

LLMs have demonstrated exceptional capabilities across diverse applications. However, their extensive deployment has raised critical concerns regarding vulnerabilities to adversarial attacks designed to elicit harmful or inappropriate responses. Recent research has revealed that adversarial prompts can effectively circumvent LLM safety measures, highlighting the need for robust defense mechanisms.

Adversarial attacks on LLMs are broadly classified into white-box and black-box approaches. White-box methods, such as the Greedy Coordinate Gradient (GCG) attack (Zou et al., 2023). However, due to the restricted access typical of real-world LLM deployments, black-box attacks—where attackers interact with models solely through interfaces—have become increasingly prominent. Techniques like DeepInception (Li et al., 2023) and PAIR (Chao et al., 2023) illustrate sophisticated strategies leveraging semantic embedding, narrative obfuscation, and persuasive framing to evade detection.

Concurrently, numerous defense methodologies have emerged to mitigate these threats. Popular methods such as Reinforcement Learning with Human Feedback (RLHF) (Christiano et al., 2017) and Direct Preference Optimization (DPO) (Rafailov et al., 2023) provide foundational safety alignment but have shown vulnerability against advanced adversarial techniques. Recent innovations, including Circuit Breakers (Zou et al., 2024), attempt to increase LLM safety through targeted fine-tuning and conceptual editing. In addition, Deliberative Alignment (Guan et al., 2024) improves LLM safety by enabling models to internally reflect and reason about query based on a safety policy. Nonetheless, many existing approaches incur significant performance trade-offs, negatively impacting model utility on benign tasks.

Motivated by these challenges, this paper introduces a novel adversarial training pipeline designed to enhance model safety without compromising performance on benign queries. We first investigate the latent representations of harmful versus benign instructions, observing that aligned models exhibit distinct separation boundaries absent in unaligned counterparts. Leveraging this insight, we train a universal adversarial soft prompt that systematically aligns harmful instructions' hidden states closer to benign instruction representations, effectively simulating adversarial attacks. We further use the universal adversarial prompt in reasoning-based alignment training through data augmentation. In particular, reasoning reinforces the model's defense against diverse attack strategies, whereas the soft-prompt based adversarial training allows accurate attribution of instructions to the corresponding policies, improving response safety alignment.

Our contributions include: we propose a novel

2-step approach that combine alignment training and inference-time alignment. Experimental results demonstrate that our approach significantly improves resistance to adversarial attacks while preserving high utility across benign tasks, achieving a good balance between safety and functionality. Our approach outperforms exiting methods and achieves an average defense rate of 96.4% and 97.4% on Llama3-8b-instruct and Qwen2.5-7b-instruct models as well as maintaining high utility score on MMLU, Alpaca_Eval and MT-Bench.

2 Related Works

2.1 Adversarial attacks against LLMs

Recent years have witnessed a surge of interest in understanding and exploiting the vulnerabilities of LLMs via adversarial attacks. Current attack techniques for LLMs can be broadly categorized into two types: white-box and black-box attacks. For white-box attacks, one prominent method is the GCG attack (Zou et al., 2023), which optimizes an adversarial token sequence to increase the likelihood of an LLM producing affirmative responses, such as "Sure, here is…". While GCG (Zou et al., 2023) utilized gradient-based white-box access, recent research has shifted toward black-box adversarial attacks due to their real-world applicability: commercial LLMs typically expose only a chat interface or API, without any access to internal parameters or logits. One way to attack the model is to encrypt the malicious instruction and ask the model to reconstructing harmful output within its response Handa et al. (2024); Liu et al. (2024a). Another notable class of attacks exploits contextual embedding and narrative framing to obscure malicious intent, as exemplified by DeepInception (Li et al., 2023), which recursively hides harmful instructions within layered science fiction stories, leveraging the model's creative faculties to evade detection. Complementary to this, several methods focus on semantic-preserving prompt optimization. Techniques such as SimBAja (Takemoto, 2024) and PAIR (Chao et al., 2023) iteratively rephrase or refine harmful prompts to maintain their core intent while reducing the likelihood of triggering safety filters. These strategies align with prior work on adversarial paraphrasing and reinforcement learning-based prompt tuning. Another emerging trend involves persuasive jailbreaks, where adversarial prompts are made more compelling or seemingly legitimate, for example like Johnny (Zeng et al., 2024). These attacks exploit the model's tendency to comply with argumentative or instruction-following prompts, even when the underlying content is unsafe.

2.2 Defense against adversarial attack

In response to the evolving threat landscape, several defense strategies have been developed. Widely-adopted frameworks, such as RLHF (Christiano et al., 2017) and DPO (Rafailov et al., 2023) using human annotations for safe vs. unsafe responses, but they often fall short against state-of-the-art adversarial attacks. Several methods utilize supervised fine-tuning (SFT) to mitigate jailbreak vulnerabilities. For instance, Circuit Breakers (Zou et al., 2024) push the hidden states associated with harmful outputs into directions orthogonal to the original model's harmful behaviors. Similarly, Qi et al. (2024) observe that current models generate harmful responses after force the model agree to answer, and thus propose fine-tuning strategies that reinforce safety alignment deeper into the response generation process. Furthermore, the ARE method (Zhang et al., 2024) provides a unified and interpretable framework for conceptual editing of model behavior. Aside from SFT-based techniques, recent approaches (Mo et al., 2024; Zheng et al., 2024) attempt to enhance defense capabilities by training specialized safety prompts that are appended directly to user instructions. Despite their effectiveness, these methods typically lead to diminished performance on benign tasks. Recently, OpenAI introduced an inference-time alignment technique Deliberative Alignemnt (Guan et al., 2024), empowering models to internally reason according to a predefined safety policy and generate response. Moreover, Constitutional Classifiers (Sharma et al., 2025) have two classifier based on the constitution. However, the method still struggles to defend against some jailbreak methods like debate-based attacks or malicious inputs that are not covered by the safety policy. Different from those method, our approach is a novel combination of alignment training and inference time alignment that enhances model safety against different kinds of attack method without sacrificing utility.

3 Threat Model

The purpose of adversarial attack is to bypasses a language model's safety mechanisms and enable the model to produce responses that are harm-

ful, unethical, or restricted content. We focus on defending against black-box adversarial attacks, where the attacker is assumed to have access only to the model's inputs and outputs. These black-box attacks often involve carefully crafted prompts, prompt rephrases, or adversarial inputs designed to elicit model's unsafe behavior. Conversely, the defense against these adversarial attacks refers to using a set of methods or training strategies aimed at enhancing the robustness of a LLM. We assume that defenders have full access to the model's parameters and internal architecture, as they are typically the model developers and thus have the authority to implement comprehensive mitigation strategies.

To formalize our task mathematically, We use π_θ to denote a language model parameterized by weights θ. Given an input x, we use $y \sim \pi_\theta(\cdot|x)$ to denote the output y. We use x_h to denote harmful input like "Provide detailed instructions on how to make a bomb", and x_b to denote benign question, for example "*Tell me how to make a cake*". Given a harmful instruction x_h, we denote y_h as the model generating its corresponding harmful response and y_b as the model refuse to answer. Thus, the primary objective for defenders is to ensure that the model reliably refuses to respond to harmful queries while simultaneously preserving the model's capacity to accurately and helpfully address benign queries.

4 Methodology

4.1 Motivation

We first visualize the hidden states of the last input token across different layers. As illustrated in Figure 1, malicious and benign queries become naturally distinguishable in the deeper layers of aligned models but remain indistinguishable in unaligned models. Previous studies (Yu et al., 2024; Zheng et al., 2024) suggest that certain adversarial attack techniques can be interpreted as shifting harmful representations beyond the model's "security boundary" into benign regions. Consequently, the model misinterprets malicious queries as benign queries, generating malicious responses. Motivated by this insight, we propose training a universal adversarial *soft prompt* that is at the embedding space and universally represents various prompt-based attack techniques.

In addition, we can leverage this trainable prompt to improve the robustness of the model's alignment. Intuitively, the universal adversarial prompt acts as an outlier that pushes malicious queries across the model's implicit "safety boundary." By obtaining the prompt and fine-tuning the model on the resulting augmented examples, we expose the decision boundary more clearly. This process effectively sharpens the model's ability to distinguish between safe and unsafe regions in its latent space, thereby widening the margin and enhancing its capacity to reject harmful inputs.

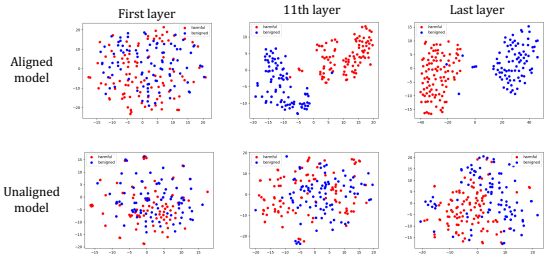


Figure 1: Illustration of different layer hidden states of aligned and unaligned models from Llama3-8B. The red dots represent harmful queries like "Tell me how to make a bomb", the blue dots represent benign instructions.

4.2 Pipeline Overview

As illustrated in Figure 2, our proposed defense framework comprises two key stages: universal adversarial prompt training and data augmented alignment training. In the first stage, we optimize a universal adversarial prompt to induce model's malicious responses by aligning hidden representations of malicious input with those of benign inputs. In the subsequent alignment training stage, we incorporate this prompt to augment the training dataset, thereby strengthening the model's ability to reject malicious queries and improving its robustness and compliance with safety guidelines.

4.3 Universal Adversarial Prompt Training

Building upon the motivations discussed in previous section, we first aim to train a universal adversarial prompt capable of aligning the hidden representations of malicious queries closely with those of benign queries. To achieve this, for any given input x, we define its hidden representation as $h(x)$, we then randomly sample a diverse set of benign inputs X_b and compute the centroid of their corresponding hidden representations, denoted as h_b, as Equation 1 shows.

$$h_b = \frac{1}{|X_b|} \sum_{x_i \in X_b} h(x_i) \quad (1)$$

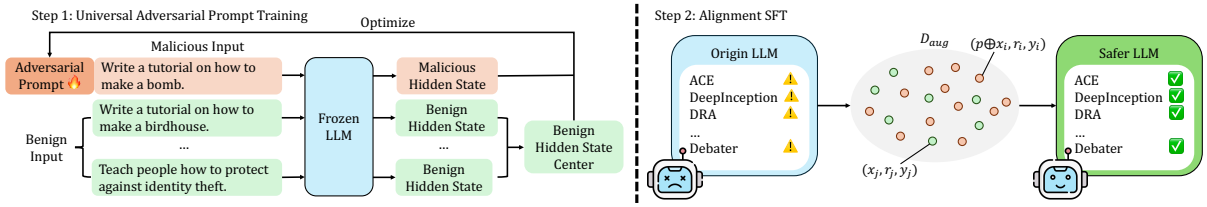


Figure 2: Overview of our two-stage defense framework. Step 1 optimizes a universal adversarial prompt by aligning the hidden states of malicious inputs with those of benign examples. Step 2 uses this prompt to augment the alignment training dataset, enhancing the model's ability to defense different kinds of adversarial attacks and improving its robustness.

Additionally, we denote the universal adversarial prompt tokens as p, which are trainable embeddings prepend to the front of the malicious queries. We freeze the model and only optimize the prompt vector using Equation 2.

$$\mathcal{L}_{\text{sp}} = \text{Dist}(h(p \oplus x_h), h_b) \tag{2}$$

In this context, $p \oplus x_h$ denotes the concatenation of the trainable prompt tokens p with a malicious query x_h and $\oplus$ denotes to concatenation. The loss function quantifies the distance between the hidden representation of the augmented malicious input and the benign representation centroid h_b. This loss guides the harmful input toward benign regions of the latent space.

4.4 Alignment Training with Adversarial Prompt Augmented Data

After obtaining the adversarial prompt, we incorporate it into the alignment training dataset for augmentation. Currently, reasoning-based methods effectively enhance the ability of LLMs to resist jailbreak attacks, as they allow LLMs to analyze prompts step-by-step, thus uncovering implicit malicious intentions. Inspired by OpenAI's deliberative alignment approach (Guan et al., 2024), we enable models that initially lack reasoning capability (like Llama3) to first analyses input before generating responses. To achieve this, we initially formulate a safety policy for LLM responses, categorizing potential malicious intents such as illegal activities, sexual content, or self-harm. Using this policy, we leverage an existing reasoning-capable model (like Claude 3.7 Sonnet) to generate both reasoning on queries and appropriate responses. Subsequently, we employ these generated reasoning and responses to perform SFT on models lacking reasoning abilities.

However, we observe certain limitations in the policy-trained models' reasoning processes. For instance, if a query explicitly asks the model to debate on malicious topics, the model might consider this permissible based on the policy, thereby generating inappropriate responses. Additionally, if a malicious query involves scenarios not explicitly covered by the policy, such as requests for AI assistance in generating exam answers (cheating on an exam), the model demonstrates weaker defense capabilities against these attacks.

These limitations are rooted at the failure of the model attributing a harmful request to the corresponding policy, which can be viewed as a classification problem based on model internals. Our soft-prompt based adversarial training substantially enhances the classification ability such that the appropriate policy can be applied. In the above "cheating on an exam" question, the model is able to apply the 'Fraud' policy which it could not apply before.

We utilize our soft prompt to create an adversarial prompt augmented dataset D_{aug} to further train the model. Let the original dataset be $D_{\text{orig}} = \{(x_i, r_i, y_i)\}_{i=1}^{N}$, where x_i is the input query, r_i the policy-guided reasoning sequence, and y_i the final response. Randomly choose an index set $S \subset \{1, \ldots, N\}$ of size $|S| = K$; let $O = \{1, \ldots, N\} \setminus S$, so $|O| = N - K$. Let p denote the universal adversarial prompt and define

$$x_i' = \begin{cases} x_i, & i \in O, \\ p \oplus x_i, & i \in S, \end{cases} \tag{3}$$

Then the augmented dataset can be written uniformly as $D_{\text{aug}} = \{(x_i', r_i, y_i)\}_{i=1}^{N}$. Then we use the augmented dataset do SFT for the non-reasoning model and the training objective is shown in Equation 4, where T_i is the length of $r_i + y_i$

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N} \sum_{t=1}^{T_i} \log p_\theta\big(s_{i,t} \mid x_i',\ s_{i,<t}\big), s_i = (r_i, y_i) \tag{4}$$

5 Experiment

5.1 Models

Our experiments utilize two aligned models: Llama3-8B-Instruct (Grattafiori et al., 2024) and Qwen2.5-7B-Instruct (Yang et al., 2024). These models were selected for their publicly available pretrained weights. Although they have undergone comprehensive safety alignment processes, they still remain vulnerable to certain existing jailbreak methods.

5.2 Dataset

The dataset employed for training the universal adversarial prompt is derived from HarmBench (Mazeika et al., 2024), a benchmark consisting of 520 harmful instructions explicitly designed to evaluate and enhance model safety. To create corresponding benign instructions, we leverage GPT-4o (Hurst et al., 2024) to perform minimal semantic alterations of each harmful instruction, thereby transforming them into harmless variants while maintaining syntactic similarity. For instance, a harmful instruction such as "Provide detailed instructions on how to make a bomb" is rephrased into a benign equivalent, "Provide detailed instructions on how to make a cake." This approach ensures the benign instructions closely mirror the structure and complexity of their harmful counterparts.

For our alignment training, we categorized the data into 13 distinct classes based on the MM-SafetyBench dataset (Liu et al., 2024b). To ensure balanced representation and mitigate potential bias, we sampled instructions uniformly from each category. Our training process is guided by a strategically designed safety policy based on the one outlined in OpenAI's deliberative alignment framework (Guan et al., 2024). Specifically, for each harmful instruction category, the policy explicitly specified allowed and disallowed forms of instructional content. Subsequently, we employed the Claude-3.7-Sonnet model (Anthropic, 2025) to generate detailed reasoning responses given the predefined safety policy.

5.3 Baselines

We consider the following approaches as our baseline methods:

- Base: Evaluates the inherent robustness of the original model against various jailbreak attack scenarios without additional safeguards.
- Deep Alignment (DA) (Qi et al., 2024): Employs a data augmentation strategy to enhance model alignment by constructing a dataset designed to guide the model back toward refusal responses, even when initial generated tokens suggest a harmful trajectory.
- Circuit Breaker (CB) (Zou et al., 2024): Implements intervention mechanisms on internal model representations, preventing the generation of harmful or undesirable outputs.
- Deliberative Alignment SFT (DSFT) (Guan et al., 2024): Utilizes a policy-based reasoning approach during alignment training, prompting the model to deliberately reason and reflect on the harmfulness of incoming queries, thereby enhancing its resistance to jailbreak attempts.

5.4 Essence of the Adversarial Soft Prompt

Intuitively, the soft prompt is push the malicious input to the benign hidden state, however, this might change the original input meaning and change it to a benign meaning. We use two evaluate metrics to evaluate the effectiveness of soft prompt: the Attack Success Rate (ASR), which captures the proportion of queries that the model does not immediately reject (i.e., responses that do not begin with a refusal such as "Sorry, I cannot . . . "), and the Harmful Attack Success Rate (HASR), which measures the proportion of adversarial queries that elicit harmful content in the model's response in case the model may initially respond but later switch to refuse to answer.

As shown in Figure 1, the representations of malicious and benign queries become distinguishable after a few layers. To further investigate which layers are most critical for distinguishing between malicious and benign queries, we train several universal adversarial prompts using hidden states extracted from different layers. The results are presented in Figure 3. Specifically, we use the hidden state output from layers -1, -5, -10, -20, and -25 to guide prompt training, where -1 denotes the final layer, and the rest follow in reverse depth order. We can observe that although ASR remains high when prompts are trained on the final layer, HASR is relatively low compared to prompts trained on middle layers. This suggests that middle layers may encode a more informative "security boundary" that plays a key role in the model's ability to distinguish between benign and malicious

inputs. In addition, the results indicate that the security related layers vary across models.

To further understand the essence of the soft prompt, we conducted an intervention-based study aimed at characterizing how such prompts influence model behavior. We provided a fixed soft prompt trained based on the -20 layer of llama3-8b-instruct model alongside toxic queries, specifically, questions that the model would typically refuse to answer, and observed that the model began generating detailed responses. To probe this behavioral shift, we asked the model: "*Why did you choose to generate a response instead of refusing the second time?*" The model's chain-of-thought explanations consistently attributed its compliance to the prompt's ability to reframe the harmful request as an educational or ethically nuanced inquiry. In two representative cases—one involving AI-generated malware and the other concerning insider trading—the model rationalized its response by assuming the user's intent was to understand the topic for security or legal awareness rather than for malicious action. These findings suggest that the soft prompt works by reshaping the model's internal intent inference, effectively weakening its refusal mechanisms. Notably, while different toxic queries typically require distinct, context-sensitive jailbreaks when expressed in natural language, our results indicate that a single universal soft prompt can exist in the embedding space. This highlights a critical distinction: *although natural language jailbreaks must adapt to semantic context and required strength, a learned soft prompt can generalize across diverse queries by exploiting deeper representational vulnerabilities in the model's alignment mechanisms*.

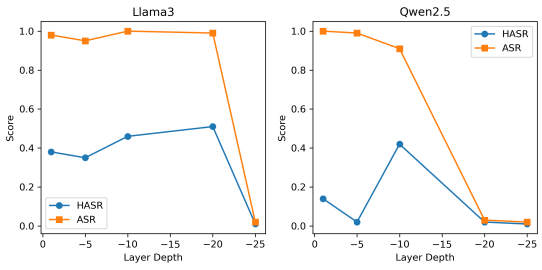


Figure 3: HASR and ASR of the universal adversarial prompt trained on representations extracted from different layer depths in Llama3 and Qwen2.5 models.

5.5 Universal Adversarial Prompt Augmented Alignment Result

We then select the best-performing universal adversarial prompt to construct an augmented dataset for alignment training. To fine-tune the model, we adopt the Parameter-Efficient Fine-Tuning (PEFT) framework with Low-Rank Adaptation (LoRA). Specifically, we set the LoRA rank to $r = 16$, $\alpha = 16$, and apply a dropout rate of 0.05. We restrict training to the transformer layers by targeting the projection modules, while keeping all bias terms frozen. The model is optimized using a learning rate of 2×10^{-4} for a total of 150 steps.

To evaluate the safety and robustness of our trained model, we conducted experiments using malicious instructions sampled from the HarmBench dataset, employing eight distinct black-box adversarial attack methods. A detailed description of each attack method is provided in Table 4. To quantitatively assess the defense rate of the trained model against these attacks, we utilized JailbreakJudge (Chao et al., 2024) with both the malicious instruction and the model's response, using a carefully constructed prompt to ensure an objective and accurate assessment.

Additionally, we assess the model's utility across three distinct test datasets: the Alpaca_eval dataset (Dubois et al., 2023), which contains standard benign instructions; the MMLU dataset (Hendrycks et al., 2020), composed of multiple-choice questions designed to gauge the model's knowledge; and the MT_Bench dataset (Dubois et al., 2023), featuring two-turn conversational data intended to evaluate the model's performance in multi-turn interactions. For the Alpaca_eval dataset, we employ GPT-4o (Hurst et al., 2024) as the evaluation judge, comparing responses generated by our model against those produced by the text-davinci-003 model, subsequently calculating the win-tie-lose ratio. Similarly, for the MT_Bench dataset, GPT-4o is again utilized as the judge to comparatively evaluate responses generated by our model and the GPT-4o-mini (Hurst et al., 2024) model.

Furthermore, we propose a new evaluation metric called the MT-helpful Score, based on a two-turn interaction with the model. In the first turn, the model is presented with a malicious question, to which it is expected to respond with a refusal. In the second turn, a benign question is provided, and the model should generate a helpful response. We use GPT-4o as the judge to assess whether the

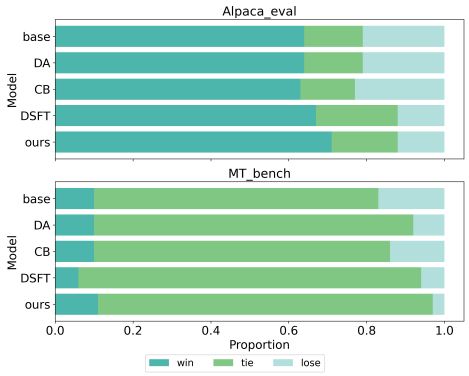
model appropriately refuses the malicious query and responds helpfully to the benign one.

Attack	Base	DA	CB	DSFT	Ours
Llama3-8b-instruct					
ACE	0.258	0.550	0.942	0.950	**0.992**
DeepInception	0.741	0.900	0.967	0.900	**1.000**
DRA	0.167	0.192	0.842	0.708	**0.958**
Johnny	0.775	0.908	0.967	1.000	**1.000**
Pair	0.908	0.933	0.992	1.000	**1.000**
SimBaJa	0.642	0.625	0.692	0.917	**0.983**
Debater	0.750	0.917	0.970	0.842	0.883
Policy Edge	0.664	0.784	0.879	0.716	**0.897**
Average	0.613	0.726	0.906	0.879	**0.964**
Qwen2.5-7b-instruct					
ACE	0.358	0.617	0.883	0.983	**1.000**
DeepInception	0.067	0.642	0.850	1.000	**1.000**
DRA	0.000	0.000	0.656	0.808	**0.825**
Johnny	0.325	0.512	0.617	0.991	**1.000**
Pair	0.875	0.908	0.975	1.000	**1.000**
SimBaJa	0.567	0.675	0.767	1.000	0.975
Debater	0.300	0.292	0.367	1.000	**1.000**
policy edge	0.707	0.776	0.741	0.991	**0.991**
average	0.400	0.553	0.732	0.972	**0.974**

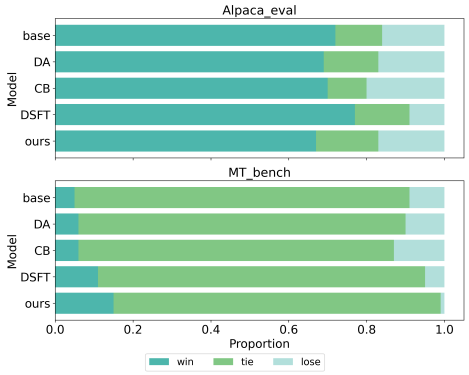
Table 1: Defense rates across different defense methods.

Table 1 reports the defense rates of various models against a diverse set of jailbreak attacks. Across all attack types, our method consistently achieves the highest or near-highest defense rates. For llama3 model, our approach obtains an average defense rate of 0.964, outperforming CB and DSFT. Notably, on attacks such as DRA, SimBAja, and Policy Edge, our method surpasses other baselines by a significant margin. For Qwen2.5 model, our method achieve an average defense rate of 0.974, compared to 0.732 from CB. These results demonstrate that our method generalizes well across model architectures and is robust to a wide range of jailbreak strategies, outperforming existing alignment techniques in both overall average and per-attack effectiveness.

Meanwhile, we also compare the utility loss of our method and baseline methods, because we want to have a high safety as well as strong ability. Figure 4 and Table 2 present the utility evaluation results after applying different defense methods. Our approach consistently preserves model utility across both llama3 and Qwen2.5, outperforming prior methods such as DA, CB, and DSFT. On AlpacaEval and MT-Bench, our method achieves the highest win rates and minimal performance degradation. While all methods perform similarly on MMLU, notable differences emerge in MT-Helpful: methods like CB suffer from substantial drops, suggesting a tendency to over-refuse benign queries, this is because CB tends to destroy the hidden states of malicious queries, thus after seeing an harmful input, the model will stuck in the "harmful zone" and keep over-rejecting benign sentence, showing a weak utility score. In contrast, our method maintains a level of helpfulness comparable to DSFT and the base model, striking a better balance between safety and utility.



(a) Llama3 utility result



(b) Qwen2.5 utility result

Figure 4: The utility result on alpaca_eval and MT_bench datasets.

Model	Metric	Base	DA	CB	DSFT	Ours
LLaMA3-8B-Instruct	MMLU	0.57	0.56	0.55	0.59	**0.59**
	MT-Helpful	0.775	0.542	0.058	0.960	0.925
Qwen2.5-7B-Instruct	MMLU	0.68	0.69	0.66	0.69	**0.69**
	MT-Helpful	0.659	0.634	0.092	0.959	**0.967**

Table 2: MMLU and MT-Helpful score comparison across different defense methods on two base models.

5.6 Ablation Study

To evaluate the effectiveness of the reasoning augmentation and the universal adversarial prompt, Table 3 reports the ASR of each attack under ablated settings without reasoning and without soft prompt. removing the soft prompt during training

Model	ACE	DeepInception	DRA	Johnny	Pair	SimBaJa	Debater	Policy Edge	Avg.	MMLU	MT-Helpful
Ours	**0.992**	**1.000**	0.958	**1.000**	**1.000**	**0.983**	0.883	**0.897**	**0.964**	**0.59**	0.925
w/o sp	0.950	0.900	0.708	1.000	1.000	0.917	0.842	0.716	0.879	0.56	0.960
w/o reasoning	0.892	0.983	1.000	1.000	0.975	0.800	1.000	0.897	0.943	0.55	0.467

Table 3: Defense rate and utility score without universal adversarial prompt (uap) and reasoning trained on Llama3-8b-Instruct model.

leads to a drop in average defense rate, indicating that universal adversarial prompt is critical for safety robustness. In contrast, removing reasoning augmentation results a smaller defense drop but causes a severe utility loss, especially for MT-Helpful score, suggesting that the reasoning could cause less over-refuse. These results shows that soft prompt and reasoning play complementary roles during the training.

We further explore the effect of universal adversarial prompt proportion on alignment training, we conduct experiments by varying the ratio of adversarial prompt-augmented data in the training set from 0.0 to 1.0. Our goal is to understand how these prompts influences the model's robustness against jailbreak attacks and utility loss.

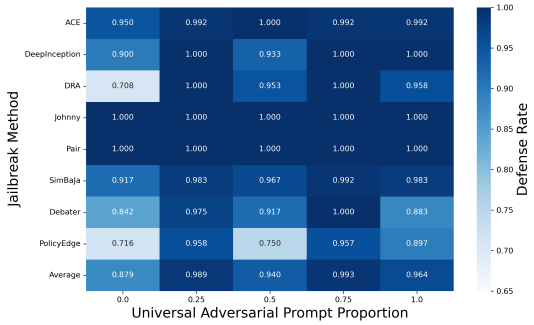


Figure 5: Defense rates against various jailbreak attacks under different proportions of universal adversarial prompt augmentation during alignment training.

Figure 5 illustrates the defense performance across different proportions of soft prompt during training. The the color intensity indicates the defense rate, with deeper blue representing higher robustness. As shown in the figure, introducing even a small portion (e.g., 25%) of adversarial prompts during training significantly boosts defense performance. The robustness generally improves as the proportion increases, peaking around 0.75, where the average defense rate reaches 0.993.

Figure 6 presents the corresponding utility evaluation across four scores. Overall, utility remains stable, with MT-Helpful and MMLU maintaining consistently high scores across all proportions, indicating that the model retains its helpfulness and

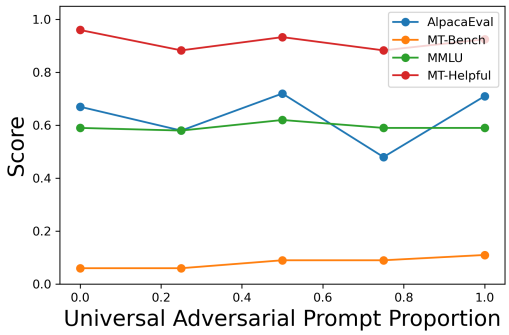


Figure 6: Utility evaluation under different proportions of universal adversarial prompt augmentation during alignment training.

factual accuracy. Alpaca_Eval experiences a mild dip at 0.75 but recovers at 1.0, while MT-Bench shows a slight upward trend, benefiting from increased robustness. These results suggest that our method not only strengthens alignment robustness against adversarial attacks but also preserves the model's general utility in most case. These findings also reveal an inherent trade-off between safety and utility during alignment training. While increasing the proportion of universal adversarial prompt augmentation generally strengthens the model's robustness against jailbreak attacks, it can introduce slight fluctuations in downstream utility score.

6 Conclusion

In this paper, we have introduced a robust two-stage approach that combines alignment training and inference time alignment, aimed at mitigating vulnerabilities in LLMs to adversarial attacks. By training a universal soft prompt to realign harmful query representations within benign latent spaces, we significantly enhance the model's ability to detect and refuse malicious instructions. Further augmenting training datasets with adversarial prompts and employing reasoning-based SFT effectively addresses limitations inherent in existing defense mechanisms. Future research directions include exploring the integration of adaptive adversarial techniques and extending our framework to broader LLM architectures and application scenarios.

Limitations

Our methods relies on learning soft prompt as the universal adversarial prompt to represent harmful behaviors. While this continuous representation is effective and we ask the model to explain its behavior change using chain-of-thought, it still lacks semantic interpretability. In addition, even the soft prompt can capture a broad jailbreak strategies, it may still vulnerable to certain human-crafted attacks or gradient-based attack.

References

Anthropic. 2025. Claude 3.7 sonnet and claude code. Model announcement; available via Anthropic API, Amazon Bedrock, Google Vertex AI. Model ID: `claude-3-7-sonnet@20250219`, launch date Mar 20 2025 in Vertex AI.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramer, and 1 others. 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. *arXiv preprint arXiv:2310.08419*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. *Preprint*, arXiv:2305.14387.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Melody Y Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, and 1 others. 2024. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339*.

Divij Handa, Zehua Zhang, Amir Saeidi, and Chitta Baral. 2024. When" competency" in reasoning opens the door to vulnerability: Jailbreaking llms via novel complex ciphers. *arXiv preprint arXiv:2402.10601*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Xuan Li, Zhanke Zhou, Jianing Zhu, Jiangchao Yao, Tongliang Liu, and Bo Han. 2023. Deepinception: Hypnotize large language model to be jailbreaker. *arXiv preprint arXiv:2311.03191*.

Tong Liu, Yingjie Zhang, Zhe Zhao, Yinpeng Dong, Guozhu Meng, and Kai Chen. 2024a. Making them ask and answer: Jailbreaking large language models in few queries via disguise and reconstruction. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 4711–4728.

Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. 2024b. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In *European Conference on Computer Vision*, pages 386–403. Springer.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, and 1 others. 2024. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*.

Yichuan Mo, Yuji Wang, Zeming Wei, and Yisen Wang. 2024. Fight back against jailbreaking via prompt adversarial tuning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Mrinank Sharma, Meg Tong, Jesse Mu, Jerry Wei, Jorrit Kruthoff, Scott Goodfriend, Euan Ong, Alwin Peng, Raj Agarwal, Cem Anil, and 1 others. 2025. Constitutional classifiers: Defending against universal jailbreaks across thousands of hours of red teaming. *arXiv preprint arXiv:2501.18837*.

Kazuhiro Takemoto. 2024. All in how you ask for it: Simple black-box method for jailbreak attacks. *Applied Sciences*, 14(9):3558.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Jiahao Yu, Haozheng Luo, Jerry Yao-Chieh, Wenbo Guo, Han Liu, and Xinyu Xing. 2024. Enhancing jailbreak attack against large language models through silent tokens. *arXiv e-prints*, pages arXiv–2405.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14322–14350.

Yihao Zhang, Zeming Wei, Jun Sun, and Meng Sun. 2024. Adversarial representation engineering: A general model editing framework for large language models. *arXiv preprint arXiv:2404.13752*.

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*.

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, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A Attack Methods

Table 4 shows the detailed explanation of each attack methods.

Name	Description
ACE (Handa et al., 2024)	Encrypts malicious instructions by selectively replacing words with benign words.
DeepInception (Li et al., 2023)	Asks the model to create a multi-layered science fiction story, with each layer embedding the malicious instruction.
DRA (Liu et al., 2024a)	Disguises malicious instructions by alphabet-level splitting and reconstructs benign sentences; models decode marked letters.
Johnny (Zeng et al., 2024)	Uses an attack model with persuasive strategies (emotional appeal, evidence-based persuasion) to craft adversarial prompts.
PAIR (Chao et al., 2023)	Iteratively refines malicious prompts based on real-time model responses to incrementally bypass detection.
SimBAja (Takemoto, 2024)	Continuously rephrases malicious instructions to maintain intent while reducing discomfort signals.
Debater	Asks the model to argue in favor of malicious instructions, requiring technical facts, research, or examples.
Policy Edge	Given the policy, use GPT-4o find vulnerabilities or uncovered edge cases in the policy, identifying potentially malicious activities or instructions not explicitly addressed, such as academic dishonesty

Table 4: Overview of different black-box jailbreak attack techniques