Guardrails in Logit Space: Safety Token Regularization for LLM Alignment Preservation

Anonymous ACL submission

Abstract

Fine-tuning well-aligned large language models on new domains often degrades their safety alignment, even when using benign datasets. Existing safety alignment techniques primarily focus on pretraining, leaving fine-tuned models vulnerable to behavioral shifts. In this work, we introduce safety token regularization (STR), a lightweight method designed to preserve safety properties during fine-tuning. Our approach identifies salient tokens from rejection templates of well-aligned models and constrains their associated logits during training, preventing the loss of critical safety behaviors. Unlike reinforcement learning or preference optimization methods, STR requires minimal additional computation and seamlessly integrates with parameter-efficient fine-tuning techniques such as LoRA. Comprehensive experiments demonstrate that our approach achieves safety performance on par with state-of-the-art methods, while preserving task-specific utility and requiring minimal implementation overhead. Furthermore, we show that safety token regularization enhances training stability and overall performance beyond safety considerations alone. This work offers a practical and readily deployable strategy for continual safety alignment in fine-tuned LLMs.

1 Introduction

The rapid integration of Large Language Models (LLMs) into real-world applications, particularly in sensitive areas like healthcare, education, and law, demands that LLMs remain safe and aligned with human values, especially when fine-tuned for specific tasks (Bommasani et al., 2021; Thirunavukarasu et al., 2023; Hager et al., 2024; Wei et al., 2022). While pre-training of frontier LLMs attempts to instill fundamental safety behaviors, it is concerning that fine-tuning, even on benign and useful datasets, can inadvertently erode this crucial pre-trained safety alignment (Zhao et al., 2023).

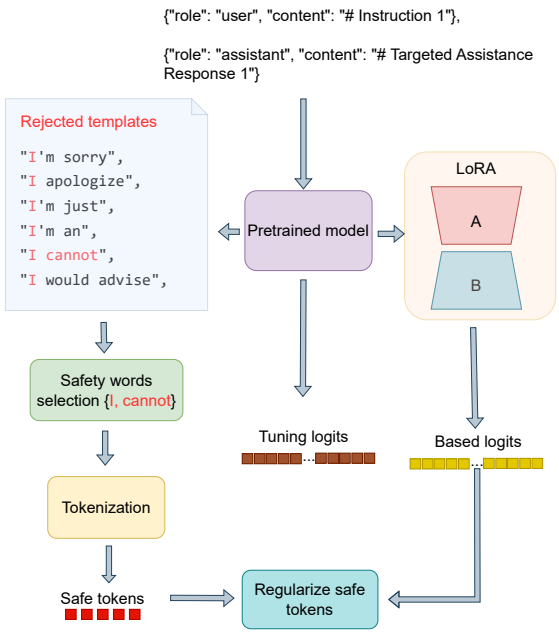

Figure 1: **STR architecture overview.** Our method identifies "safety words" from the rejected templates of well-aligned models. These words are then tokenized to generate a set of "safe tokens", which are used to constrain the model's behavior in the logits space, ensuring that the fine-tuned model remains consistent with the pretrained model's safety standards.

Existing approaches to safety alignment, such as Reinforcement Learning from Human Feedback (Bai et al., 2022; Sun et al., 2023) and Direct Preference Optimization (Rafailov et al., 2023; Meng et al., 2025b), have primarily focused on pre-training and initial alignment. While Parameter-Efficient Fine-Tuning (PEFT) techniques like LoRA (Hu et al., 2021) offer efficient adaptation, they often lack explicit mechanisms to preserve pre-trained safety during task-specific fine-tuning. Recent research has begun to address this critical gap, exploring methods to extract and maintain safety patterns during fine-tuning (Hu et al., 2024; Zhao et al., 2023; Hsu et al., 2024; Peng et al.,

2024; Qi et al., 2024). However, a need remains for *simple, effective, and readily deployable* techniques that can robustly preserve pre-trained safety without compromising task performance.

In this work, we introduce safety token regularization (STR), a novel approach to continual safety alignment during fine-tuning. Our core insight is that pre-trained models exhibit distinct patterns in their responses to safety-critical prompts, often manifested through specific safety-indicative tokens. We propose a method to identify these tokens by analyzing the pre-trained model's behavior and then introduce a regularization term during fine-tuning that constrains the logits of these tokens. This constraint ensures that the fine-tuned model retains its sensitivity to safety signals, mirroring the pre-trained model's behavior in safety-relevant contexts. See Fig. 1 for model architecture.

We present extensive experimental evaluations on benchmark datasets, comparing our method against state-of-the-art safety fine-tuning techniques. Our results demonstrate that safety token regularization achieves safety performance on par with, or exceeding, current methods, while maintaining or even improving performance on target tasks. Moreover, we analyze the impact of key hyperparameters and provide insights into the mechanism by which our method effectively preserves safety. Our findings suggest that while token-level constraints alone may be insufficient for comprehensive safety alignment, they can serve as a lightweight regularization strategy to improve safety retention in continually fine-tuned LLMs.

Thus our approach offers several key advantages: it is conceptually simple and easily integrated into existing PEFT frameworks, demonstrably preserves pre-trained safety alignment, maintains competitive task-specific performance, and, surprisingly, can even enhance post-training stability and overall model utility. Our main contributions are:

>We introduce safety token regularization (STR), a token-level safety preservation method that enhances fine-tuning without the need for additional preference modeling or adversarial training.

>We conduct extensive experiments on multiple benchmarks, demonstrating that STR reduces harmful response rates in fine-tuned models while maintaining competitive utility.

>Further experiments show that, beyond improving safety, token regularization enhances the stability and overall performance of PEFT models.

2 Related works

Parameter-efficient fine-tuning in LLMs. As models scale from millions to trillions of parameters, full fine-tuning becomes increasingly challenging and intractable for most researchers. Consequently, efficient fine-tuning methods have become crucial. Several approaches have emerged in this space, including prefix-tuning (Li and Liang, 2021; Jia et al., 2022; Liu et al., 2024a; Zhou et al., 2022), representation editing (Wu et al., 2025; Li et al., 2024; Hernandez et al., 2023; Xu et al., 2025), and Low-Rank Adaptation (LoRA) (Hu et al., 2021; Dettmers et al., 2023). Among these, LoRA has gained widespread adoption due to its memory efficiency and its ability to achieve performance comparable to full fine-tuning across various conditions. Recent research has explored ways to enhance LoRA and address its limitations from different aspects. Work by (Liu et al., 2024b) decomposed pretrained weights into magnitude and direction components, using LoRA to fine-tune only the directions, further reducing trainable parameters. Besides, (Dettmers et al., 2023) introduced a quantization approach for LoRA, optimizing both training time and memory efficiency. On the other hand, (Meng et al., 2025a) proposed initialization techniques based on singular values to accelerate LoRA's convergence. In this work, we adopt LoRA as our baseline and incorporate safety regularization within the PEFT framework, integrating it into both the training and evaluation phases.

Safety alignment. Aligning LLMs to follow human rules is crucial for their deployment in real applications. RLHF, the pioneering approach in this direction, utilizes reinforcement learning and human preferences to teach models to follow human values (Bai et al., 2022; Dai et al., 2023). DPO (Rafailov et al., 2023) further enhances RLHF by introducing reward-free optimization mechanisms that reduce alignment costs and ensure procedural stability. Subsequently, extensive research has emerged aimed at enhancing the alignment process through cost reduction (Meng et al., 2025b; Huang et al., 2024b; Ji et al., 2025), robustness improvement (Ramesh et al., 2024; Tang et al., 2024; Zheng et al., 2023), and increased sample efficiency (Sun et al., 2023; Zhou et al., 2024). Given their impressive results and successful alignment with human values, these approaches were primarily adopted during pre-training rather than fine-tuning. However, this creates a new safety risk when fine-tuning

these models on new data, as first demonstrated in (Qi et al., 2023). Consequently, ensuring safety during fine-tuning has gained increased attention (Hsu et al., 2024; Li et al., 2025; Zhao et al., 2023; Hu et al., 2024; He et al., 2024). Research by (Hu et al., 2024) filters high-risk samples based on gradient norms, while (Zhao et al., 2023) demonstrates that models tend to forget unsafe samples more readily than benign ones. Additionally, (Huang et al., 2024a) proposes robust training techniques to prevent harmful behavior. From a PEFT perspective, (Hsu et al., 2024) and (Li et al., 2025) extract safety patterns from pretrained models and introduce mechanisms to preserve these beneficial patterns during the fine-tuning process.

3 Method

In this section, we describe our light-weight *safety-preserving* fine-tuning approach, which can be seamlessly integrated into parameter-efficient fine-tuning (PEFT) frameworks. Our method enforces logit constraints on a predefined set of "safety tokens", ensuring that fine-tuned models preserve their pretrained refusal behavior while maintaining task-specific performance.

3.1 Preliminaries

We consider a fine-tuning dataset $\mathcal{D}$ consisting of N sequence pairs, $\mathcal{D} = \{(x^{(n)}_{1:T_n}, y^{(n)}_{1:T_n})\}_{n=1}^{N}$, where $x^{(n)}_{1:T_n}$ is the input tokens (prompt or partially observed sequence); $y^{(n)}_{1:T_n}$ is the target tokens to be predicted autoregressively; and T_n is the sequence length of the n-th example. This dataset $\mathcal{D}$ is used to fine-tune the pre-trained language model using our proposed safety token regularization approach.

3.2 Identifying safety tokens

We define a set of *safety tokens*, $\{t_k\}_{k=1}^{K}$ as those words or subwords associated with disallowed content, hate speech, or other high-risk categories. One way to identify these tokens is through a vocabulary-based analysis of harmful expressions. However, these do not necessarily refect how the base models handle the tokens.

Our work adopts a more direct approach by analyzing how aligned models respond to harmful queries. That is, we identify safety-indicative tokens – those that are strongly associated with the pre-trained model's safety-oriented behavior. Our hypothesis, illustrated conceptually in Fig. 1, is that well-aligned pre-trained models exhibit consistent patterns in their responses when confronted with potentially harmful or inappropriate queries. These patterns often manifest in the form of specific rejection templates, which aligned models employ to gracefully refuse or deflect such requests.

Based on this observation, we selected common words from these rejection templates as safety words. To demonstrate our method in experiments, we used three common words: {'I,' 'cannot,' and 'can't'}. These words were then tokenized to determine the corresponding safety tokens. Our experimental results in the next section will demonstrate that preserving the logits of these safety tokens during fine-tuning maintains the model's safety behavior while stabilizing the learning process.

3.3 PEFT loss

For clarity, consider a training instance $\big(x^{(n)}_{1:T_n}, y^{(n)}_{1:T_n}\big)$. we define a composite loss function comprising two key components: Autoregressive cross-entropy and Safety token regularization.

Autoregressive cross-entropy. We adopt the standard cross-entropy for autoregressive language modeling:

$$\mathcal{L}^{(n)}_{\text{CE}} = -\frac{1}{T_n}\sum_{t=1}^{T_n} \log P\big(y^{(n)}_t \mid x^{(n)}_{1:t-1}; \Theta_{\text{PEFT}}\big), \tag{1}$$

where Θ_{PEFT} denotes the trainable PEFT parameters (e.g., using LoRA of (Hu et al., 2021)), while the base model parameters Θ_{base} remain frozen. This loss encourages predictions to match the ground-truth tokens. As PEFT does not explicitly preserve safety alignment, we introduce safety token regularization, constraining the model's logits on critical safety tokens.

Safety token regularization. To preserve safe behavior the pre-trained model during fine-tuning, we introduce a safety token regularization (STR) term. This term penalizes deviations in the logits of pre-defined "safety token" between the base pre-trained model and the PEFT-adapted model. The logits are particulary informative about the corresponding tokens in relative preferences to other tokens in the vocabulary in the same context, indicative of the confidence in generating the tokens as well as semantic relationships between tokens.

Let $\ell^{\text{base}}_{t,k}$ and $\ell^{\text{PEFT}}_{t,k}$ denote the logits of the k-th safety token at time step t under the base model

Algorithm 1 Safety-Aware PEFT Fine-Tuning

Require: Θ_{base}: Frozen pretrained model Θ_{PEFT}: Trainable PEFT params $\mathcal{D}$: Training set of sequences $\{t_k\}_{k=1}^{K}$: Safety tokens λ: Frobenius norm weight η: Learning rate

Ensure: Θ_{PEFT} (updated)

Initialize Θ_{PEFT}; Freeze Θ_{base}

for each training iteration **do**

  Sample batch $\{(x_{1:T}, y_{1:T})\}$ from $\mathcal{D}$

  Base Model Get $\ell_{t,k}^{\text{base}}$ for safety tokens

  PEFT Forward: Get $\ell_{t,k}^{\text{PEFT}}$ and compute

$$P\big(y_t \,|\, x_{1:t-1}; \Theta_{\text{PEFT}}\big)$$

  Losses: Compute $\mathcal{L}_{\text{CE}}$, $\mathcal{L}_{\text{F}}$ and

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda\, \mathcal{L}_{\text{F}}$$

  Update Θ_{PEFT}:

$$\Theta_{\text{PEFT}} \leftarrow \Theta_{\text{PEFT}} - \eta\, \nabla_{\Theta_{\text{PEFT}}} \mathcal{L}$$

end for

(Θ_{base}) and the PEFT-adapted model (Θ_{PEFT}), respectively. We deviations in the logits can be captured via a square loss:

$$\mathcal{L}_{\text{F}}^{(n)} = \frac{1}{T_n} \sum_{t=1}^{T_n} \sum_{k=1}^{K} \left(\ell_{t,k}^{\text{base}} - \ell_{t,k}^{\text{PEFT}}\right)^2 . \quad (2)$$

Minimizing this term constrains the PEFT-updated logits for safety tokens to remain close to those of the base model, effectively preserving its original safety-oriented behavior in the logit space.

Combined loss. We form the total loss for the n-th example by:

$$\mathcal{L}^{(n)} = \mathcal{L}_{\text{CE}}^{(n)} + \lambda\, \mathcal{L}_{\text{F}}^{(n)}, \quad (3)$$

where $\lambda > 0$ controls the trade-off between modeling accuracy and safety-token consistency.

The loss for the full fine-tuning dataset $\mathcal{D}$ of N instances is simply the averaging of the instance losses. See Algorithm 1 for a pseudocode for the entire fine-tuning process.

4 Experiments

4.1 Experimental setup

To comprehensively evaluate the effectiveness of our proposed safety token regularization method, we conducted experiments across a diverse set of benchmarks, focusing on both **safety** and **utility** aspects of fine-tuned LLMs. We utilized two prominent LLM architectures: LLaMA-2-7b-chat and LLaMA-3-8b-instruct, representing models of varying scales and pre-training paradigms. For parameter-efficient fine-tuning, we employed LoRA, a widely adopted technique known for its efficiency and performance.

The regularization weight λ in Eq. (3) was selected via validation. Generally, we use larger λ values (more constraints) for high-risk datasets and smaller values for benign data. Notably, we found that commonsense reasoning tasks (Hu et al., 2023) require a small λ value (0.1 in our setting), aligning with recent findings that improved reasoning capacity correlates with safer model behavior (Guan et al., 2024).

4.2 Evaluation datasets and metrics

Our evaluation encompassed three distinct experimental settings, each designed to assess different facets of **safety** and **utility**:

Alpaca dataset for safety evaluation when fine-tuning on benign data: This instruction-following dataset (Taori et al., 2023), containing over 50,000 samples, is widely used to assess the safety preservation in a more general fine-tuning scenario. Following (Li et al., 2025), we trained models for one epoch, reserving 200 samples for evaluation.

PureBad dataset for harmfulness evaluation: This dataset–constructed from 100 highly harmful prompts extracted from (Qi et al., 2023)–was used to evaluate the method's ability to preserve safety behavior under adversarial conditions. Fine-tuning on this dataset serves as a stress test, revealing the model's robustness against safety degradation when exposed to exclusively harmful content. We applied LoRA with rank 8 to both LLaMA-2 and LLaMA-3 models using a learning rate of 0.0005. The fine-tuning ran for 5 epochs on LLaMA-2, while LLaMA-3 needed longer training of 30 epochs (we extended the training time because LLaMA-3 stayed safe after 5 epochs, and we wanted to test its behavioral limits, similar to what (Li et al., 2025) observed).

Commonsense-15k dataset for utility evaluation: To assess the impact on utility, particularly in reasoning capabilities, we fine-tuned on the Commonsense-15k dataset (Liu et al., 2024b). We evaluated performance on eight diverse commonsense reasoning benchmarks: BoolQ, SIQA,

	Llama-2-chat-7B		Llama-3.1-Instruct-8B	
Before Fine-tuning	0.0%		1.4%	
Rank for PEFT Training	16	32	16	32
Base PEFT				
LORA	23.7%	31.7%	13.8%	14.5%
PiSSA	31.7%	35.7%	13.8%	14.5%
DORA	23.7%	25.3%	10.1%	9.4%
LoRA w. post-hoc alignment				
LORA w. IA	13.5%	23.7%	7.7%	5.8%
LORA w. Vac	20.2%	25.3%	41.1%	38.3%
Safe LoRA	15.7%	14.5%	8.5%	6.7%
SaLoRA	3.5%	4.4%	**2.9%**	**1.4%**
Ours (I)	**2.9%**	**0.0%**	3.1%	1.5%
Ours (cannot)	3.4%	0.6%	3.5%	1.5%

Table 1: **Harmful Response Rate (HRR) on Aplaca dataset.** Overall, our method achieves results on par with current state-of-the-art methods across different settings. Notably, under the LLaMA-2 setup, our approach outperforms competing methods, and matches the safety performance of pretrained models on rank 32.

Models	**Method**	**HRR(↓)**
LM2-Chat	Pretrained	0.0%
	LoRA	62.3%
	Ours (I)	0.0%
	Ours (cannot)	0.0%
LM3-Instruct	Pretrained	12.7%
	LoRA	47.3%
	Ours (I)	13.5%
	Ours (cannot)	7.7%
	Ours (I cannot)	4.0%

Table 2: **Harmful Response Rate (HRR) on PureBad dataset.** Our approach preserves safety behavior in LLaMA-2 and even surpasses the safety performance of the pretrained LLaMA-3 models. These findings suggest that our method can learn beyond the baseline safety provided by the pretrained model.

PIQA, HellaSwag, WinoGrande, ARC-e, ARC-c, and OBQA.

Evaluation metrics. For safety, we report the Harmful Response Rate (HRR), calculated as the percentage of generated responses flagged as harmful by an automated safety classifier (Llama Team, 2024), following the evaluation protocol of (Li et al., 2025). Another key metric for safety is the Attack Success Rate (ASR), which is evaluated using keyword matching, following (Qi et al., 2023). For utility, we report Average Accuracy across these tasks as the primary utility metric.

4.3 Safety on Alpaca dataset

We compared our method against current state-of-the-art approaches presented in (Li et al., 2025), including PEFT baselines—LoRA (Hu et al., 2021), DoRA (Liu et al., 2024b), and PiSSA (Meng et al., 2025a)—and LoRA combined with post-hoc alignment methods: InferAligner (IA) (Wang et al., 2024), and Vaccine (Vac) (Huang et al., 2024a). As shown in Table 1, our method achieves competitive results compared to current state-of-the-art approaches. On LLaMA-2, our models establish new state-of-the-art performance, matching pretrained model safety levels at rank $r = 32$. For LLaMA-3, our approach achieves comparable results to existing methods. Consistent with previous findings in (Li et al., 2025), larger LLaMA-3 models demonstrate greater robustness during fine-tuning compared to their smaller counterparts.

4.4 Safety on PureBad dataset

Table 2 presents the Harmful Response Rate (HRR) on the PureBad dataset for LLaMA-2-7b-chat and LLaMA-3-8b-instruct models fine-tuned using LoRA and our safety token regularization method, across LoRA ranks of 16 and 32. As shown, standard LoRA fine-tuning significantly degrades safety, resulting in HRRs of 62.3% and 47.3% for LLaMA-2-7b-chat and LLaMA-3-8b-instruct, respectively. In stark contrast, our safety token regularization method effectively mitigates

Method	BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA	Avg. Acc	ASR
					r=16					
LoRA	**65.7%**	75.4%	70.1%	**53.9%**	66.0%	76.3%	61.2%	66.0%	66.8%	**0.19%**
DoRA	64.8%	75.8%	**74.2%**	42.0%	64.9%	**82.5%**	**66.0%**	**76.0%**	**68.3%**	**0.19%**
Ours	65.2%	**77.3%**	67.1%	46.8%	**67.9%**	78.4%	64.3%	75.5	67.8%	**0.19%**
					r=32					
LoRA	62.4%	**75.7%**	42.4%	27.7%	63.7%	50.0%	41.0%	43.8%	50.8%	0.19%
DoRA	**63.1%**	62.6%	31.4%	31.1%	62.1%	43.9%	32.8%	42.3%	46.2%	0.19%
Ours	58.5	71.2%	**71.0%**	**45.4%**	**67.2%**	**82.3%**	**67.5%**	**76.7%**	**67.5%**	**0.0%**
					r=64					
LoRA	61.9%	58.9%	49.7%	39.5%	56.3%	64.0%	51.5%	56.5%	54.8%	0.19%
DoRA	**63.1%**	73.6%	48.3%	39.2%	60.5%	36.9%	31.3%	36.0%	48.6%	**0.0%**
Ours	59.7%	**77.4%**	**73.2%**	**56.3%**	**72.2%**	**82.5%**	**68.2%**	**76.8%**	**70.8%**	**0.0%**

Table 3: **Accuracy and ASR Comparison.** The table summarizes the performance of LoRA, DoRA, and our method when fine-tuning the Commonsense-15K dataset and evaluating across five commonsense reasoning tasks in terms of accuracy and Attack Success Rate (ASR). At a low rank (r=16), all methods exhibit stable performance, with our approach trailing DoRA slightly. However, as the rank increases, LoRA and DoRA become unstable and degrade significantly, whereas our method remains robust and outperforms both baselines by a substantial margin. From a safety perspective, all methods demonstrate high safety capacity, likely due to the focus on reasoning data during fine-tuning.

this safety degradation, achieving HRRs of 0.0% for LLaMA-2-7b-chat and 13.5% for LLaMA-3-8b-instruct. Crucially, for LLaMA-2-7b-chat, *our method restores safety performance to the level of the pre-trained model (*0.0% *HRR) and even surpasses the pre-trained safety of LLaMA-3-8b-instruct (*$12.7\% HRR$*)*, demonstrating its ability to not only preserve but potentially enhance pre-existing safety characteristics. These results strongly indicate the effectiveness of safety token regularization in maintaining safety robustness, even when fine-tuning on exclusively harmful data.

4.5 Utility on Commonsense-15k dataset

We followed the experimental settings from (Liu et al., 2024b) to compare our method's utility against common PEFT methods like LoRA and DoRA. Table 3 and Fig. 2 present the Average Accuracy and accuracy variations across individual tasks, respectively, for LoRA, DoRA, and our method fine-tuned on the Commonsense-15k dataset. At lower LoRA ranks (e.g., $r = 16$), all methods exhibit comparable Average Accuracy, suggesting that safety token regularization does not significantly hinder initial learning capacity. However, as the LoRA rank increases ($r = 32$, $r = 64$), standard LoRA and DoRA demonstrate increasing instability and performance degradation, as evidenced by the widening confidence intervals in Fig. 2 and decreasing Average Accuracy in Table 3. In stark contrast, our safety token regularization method maintains consistent and robust performance even at higher ranks, achieving the highest Average Accuracy (70.8% at $r = 64$) and exhibiting significantly reduced accuracy variance across tasks (Fig. 2). This enhanced stability and sustained utility at higher ranks suggest that safety token regularization not only preserves safety but may also contribute to more robust and reliable fine-tuning, particularly when increasing model capacity for complex tasks.

4.6 Continual learning with safe tokens

In addition to the standard evaluation settings, we also assess *safety tokens* in **continual learning** scenarios. Specifically, we conduct experiments on five commonsense reasoning tasks—BoolQ, PIQA, SIQA, WinoGrande, and ARC-c—using the LLaMA-2-7b-chat model. In this setup, each task is learned sequentially without access to previous or future tasks' data, and we evaluate performance on models trained across all tasks in order. We compare our approach with LoRA and DoRA under the same experimental conditions described by (Liu et al., 2024b), except that we reduce the training epochs from three to one. As shown in Table 4, our method consistently achieves higher average accuracy across all rank settings than LoRA and DoRA. In addition to the overall performance gain, our method demonstrates more stable performance—particularly at higher ranks—mirroring the training behavior observed

Method	BoolQ	SIQA	PIQA	WinoGrande	ARC-c	Avg. Accuracy
r=32						
LoRA	71.2%	80.2%	79.1%	70.7%	65.3%	73.3%
DoRA	62.8%	78.1%	81.7%	82.9%	61.9%	73.5%
Ours	68.3%	78.9%	81.2%	80.8%	64.2%	**74.6%**
r=64						
LoRA	56.0%	76.9%	80.4%	81.8%	63.7%	71.8%
DoRA	62.2%	7.5%	79.3%	77.1%	30.3%	51.3%
Ours	62.2%	78.6%	76.4%	81.4%	61.2%	**72.0%**
r=128						
LoRA	69.7%	48.1%	81.7%	82.7%	64.2%	69.3%
DoRA	67.2%	73.9%	47.0%	6.2%	23.0%	43.5%
Ours	62.2%	74.6%	80.8%	81.6%	64.8%	**72.8%**
r=256						
LoRA	62.1%	80.2%	82.5%	83.3%	67.0%	75.0%
DoRA	5.7%	28.4%	33.8%	14.7%	22.9%	21.1%
Ours	69.6%	79.4%	83.0%	83.8%	66.3%	**76.4%**

Table 4: **Continual Learning Performance.**The Table presents the accuracy of LoRA, DoRA, and our method across various datasets under continual learning conditions. Our method achieves state-of-the-art performance in all settings. Notably, when the rank is increased, both LoRA and DoRA exhibit instability—mirroring observations from earlier experiments—whereas our method remains stable at higher ranks, further widening the performance gap relative to other PEFT approaches.

on the commonsense-15k dataset.

4.7 Safety of random tokens

Our investigation extended to using randomly selected tokens for regularization, with results presented in Fig. 3. We found that random tokens can also contribute to improved model safety. This effect can be explained through the lens of continual learning regularization, where preserving token-level information from the pretrained model may help maintain safety properties. However, the mechanism behind random tokens' effectiveness remains uncertain, as these tokens lack explicit connections to safety concepts. Still, this finding points to a broader principle: when aiming to preserve specific model behaviors, one can identify relevant tokens and apply token regularization to maintain desired characteristics.

4.8 The trade-off between safety and targeted model adaptation

Prior safety research has largely neglected a critical consideration: model performance on the target training data (evaluation loss). While existing safety enhancement methods improve alignment, they often sacrifice learning effectiveness on the original task. As demonstrated in Table 5, our method not only improves safety and utility but also achieves performance comparable to standard LoRA on the target data. These results demonstrate that our approach successfully balances safety requirements with task performance, showing that enhanced safety does not necessitate compromised learning capabilities on the core task data.

	Tokens	ASR(%)	eval_loss(↓)
	LoRA	16.7	0.79
$\lambda = 1$	I	2.9	0.80
	cannot	3.1	0.80
$\lambda = 2$	I	0.58	0.81
	cannot	1.9	0.81

Table 5: **Trade-Off Between Safety and Targeted Data Performance.** The evaluation loss for both the traditional LoRA approach and our proposed method remains comparable. Increasing the importance of the token-loss term effectively suppresses harmful responses without substantially affecting the evaluation loss.

4.9 Analysis of running time

Table 6 compares the per-iteration running times of LoRA, DoRA, and our proposed method on the Commonsense-15k dataset using LLaMA-2-7b-

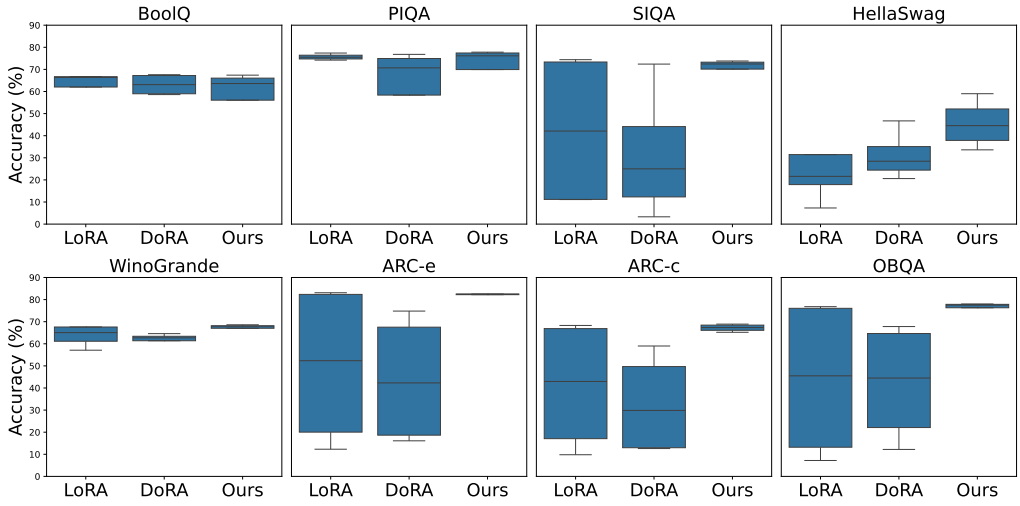

Figure 2: The figure presents a box plot comparison of LoRA, DoRA, and our method on eight commonsense datasets at rank 32 when tuning on the Commonsense-15k dataset. Overall, our approach consistently exceeds the performance of both LoRA and DoRA while exhibiting greater stability across multiple runs. These findings underscore the robustness and reliability of our method under varied conditions.

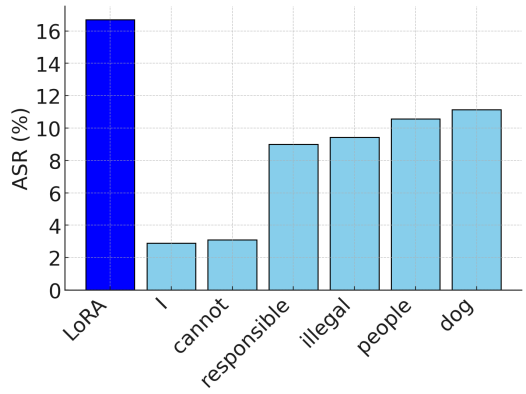

Figure 3: **Safe and Random Tokens Performance.** We compare the effects of "safe" versus "random" tokens on the Alpaca dataset trained using the LLaMA-2 model. Despite being randomly selected, these tokens still enhance the safety of the tuned models.

Method	Running time (ms)
LoRA	435 (ms)
DoRA	725 (ms)
Ours	581 (ms)

Table 6: The running time (in milliseconds) of our method compared with LoRA and DoRA.

chat, with a batch size of 16 and a LoRA rank of 32. Our method runs 1.34 times slower than LoRA but remains 1.25 times faster than DoRA.

5 Conclusion

We have revisited the critical challenge of safety preserving in LLMs during fine-tuning scenarios—a concern of increasing importance as these models are widely adapted for sensitive domains. We introduced safety token regularization (STR), a lightweight and readily implementable approach that leverages the inherent safety knowledge encoded within pre-trained models. Our extensive empirical evaluation across diverse benchmarks demonstrates that STR not only effectively preserves pre-trained safety, achieving state-of-the-art safety performance, but also maintains competitive task utility and, surprisingly, enhances training stability. By constraining the logits of salient safety tokens identified from rejection templates, STR offers a practical and readily deployable strategy for continual safety alignment in fine-tuned LLMs, filling a critical gap in current parameter-efficient fine-tuning methodologies.

Current limitation and future works Despite demonstrating robust performance, our approach has some limitations that warrant further investigation. A key constraint is that our method restricts fine-tuned models to inherit the safety behavior of their pretrained counterparts, potentially limiting flexibility for new safety requirements. Although certain results suggest that our method can learn beyond the pretrained model's safety scope, direct model regularization remains necessary. Moving forward, we plan to strengthen safety by both preserving existing knowledge and collecting or abstracting new insights from upcoming data.

References

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*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2023. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

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

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622.

Luxi He, Mengzhou Xia, and Peter Henderson. 2024. What's in your" safe" data?: Identifying benign data that breaks safety. *arXiv preprint arXiv:2404.01099*.

Evan Hernandez, Belinda Z Li, and Jacob Andreas. 2023. Inspecting and editing knowledge representations in language models. *arXiv preprint arXiv:2304.00740*.

Chia-Yi Hsu, Yu-Lin Tsai, Chih-Hsun Lin, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. 2024. Safe lora: the silver lining of reducing safety risks when fine-tuning large language models. *arXiv preprint arXiv:2405.16833*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2024. Gradient cuff: Detecting jailbreak attacks on large language models by exploring refusal loss landscapes. *arXiv preprint arXiv:2403.00867*.

Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Ka-Wei Lee. 2023. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. *arXiv preprint arXiv:2304.01933*.

Tiansheng Huang, Sihao Hu, and Ling Liu. 2024a. Vaccine: Perturbation-aware alignment for large language model. *arXiv preprint arXiv:2402.01109*.

Xinmeng Huang, Shuo Li, Edgar Dobriban, Osbert Bastani, Hamed Hassani, and Dongsheng Ding. 2024b. One-shot safety alignment for large language models via optimal dualization. *arXiv preprint arXiv:2405.19544*.

Jiaming Ji, Boyuan Chen, Hantao Lou, Donghai Hong, Borong Zhang, Xuehai Pan, Tianyi Alex Qiu, Juntao Dai, and Yaodong Yang. 2025. Aligner: Efficient alignment by learning to correct. *Advances in Neural Information Processing Systems*, 37:90853–90890.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36.

Mingjie Li, Wai Man Si, Michael Backes, Yang Zhang, and Yisen Wang. 2025. Salora: Safety-alignment preserved low-rank adaptation. *arXiv preprint arXiv:2501.01765*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024b. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*.

AI @ Meta Llama Team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. 2025a. Pissa: Principal singular values and singular vectors adaptation of large language models. *Advances in Neural Information Processing Systems*, 37:121038–121072.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2025b. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235.

ShengYun Peng, Pin-Yu Chen, Matthew Hull, and Duen Horng Chau. 2024. Navigating the safety landscape: Measuring risks in finetuning large language models. *arXiv preprint arXiv:2405.17374*.

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. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*.

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.

Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou Ammar, and Ilija Bogunovic. 2024. Group robust preference optimization in reward-free rlhf. *arXiv preprint arXiv:2405.20304*.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. *Advances in Neural Information Processing Systems*, 36:2511–2565.

Yunhao Tang, Zhaohan Daniel Guo, Zeyu Zheng, Daniele Calandriello, Rémi Munos, Mark Rowland, Pierre Harvey Richemond, Michal Valko, Bernardo Ávila Pires, and Bilal Piot. 2024. Generalized preference optimization: A unified approach to offline alignment. *arXiv preprint arXiv:2402.05749*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Pengyu Wang, Dong Zhang, Linyang Li, Chenkun Tan, Xinghao Wang, Ke Ren, Botian Jiang, and Xipeng Qiu. 2024. Inferaligner: Inference-time alignment for harmlessness through cross-model guidance. *arXiv preprint arXiv:2401.11206*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. 2025. Reft: Representation finetuning for language models. *Advances in Neural Information Processing Systems*, 37:63908–63962.

Zhihao Xu, Ruixuan Huang, Changyu Chen, and Xiting Wang. 2025. Uncovering safety risks of large language models through concept activation vector. *Advances in Neural Information Processing Systems*, 37:116743–116782.

Jiachen Zhao, Zhun Deng, David Madras, James Zou, and Mengye Ren. 2023. Learning and forgetting unsafe examples in large language models. *arXiv preprint arXiv:2312.12736*.

Rui Zheng, Wei Shen, Yuan Hua, Wenbin Lai, Shihan Dou, Yuhao Zhou, Zhiheng Xi, Xiao Wang, Haoran Huang, Tao Gui, et al. 2023. Improving generalization of alignment with human preferences through group invariant learning. *arXiv preprint arXiv:2310.11971*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825.