

# Safeguarding LLM Fine-tuning via Push-Pull Distributional Alignment

Haozhong Wang<sup>1\*</sup>, Zhuo Li<sup>2\*</sup>, Yibo Yang<sup>3,4</sup>, He Zhao<sup>5,6</sup>, Hongyuan Zha<sup>2</sup>, Dandan Guo<sup>1,3†</sup>

<sup>1</sup>School of Artificial Intelligence, Jilin University,

<sup>2</sup>The Chinese University of Hong Kong, Shenzhen

<sup>3</sup>King Abdullah University of Science and Technology,

<sup>4</sup>University of Oxford, <sup>5</sup>CSIRO's Data61,

<sup>6</sup>Faculty of Information Technology, Monash University,  
guodandan@jlu.edu.cn

## Abstract

The inherent safety alignment of Large Language Models (LLMs) is prone to erosion during fine-tuning, even when using seemingly innocuous datasets. While existing defenses attempt to mitigate this via data selection, they typically rely on heuristic, instance-level assessments that neglect the global geometry of the data distribution and fail to explicitly repel harmful patterns. To address this, we introduce Safety Optimal Transport (SOT), a novel framework that reframes safe fine-tuning from an instance-level filtering challenge to a distribution-level alignment task grounded in Optimal Transport (OT). At its core is a dual-reference “push-pull” weight-learning mechanism: SOT optimizes sample importance by actively pulling the downstream distribution towards a trusted safe anchor while simultaneously pushing it away from a general harmful reference. This establishes a robust geometric safety boundary that effectively purifies the training data. Extensive experiments across diverse model families and domains demonstrate that SOT significantly enhances model safety while maintaining competitive downstream performance, achieving a superior safety-utility trade-off compared to baselines. Our code is available at <https://github.com/kasaer/SOT>.

## 1 Introduction

As large language models (LLMs) become foundational in various industries, fine-tuning is the standard method for adapting them to new tasks. Before such fine-tuning, foundation LLMs are typically safety-aligned (e.g., RLHF (Ouyang et al., 2022), DPO (Rafailov et al., 2023)) to ensure outputs follow safety preferences and reliably refuse harmful requests. However, recent studies reveal

that this fine-tuning process presents a critical vulnerability in model safety (Yang et al., 2023b; Wei et al., 2023; Qi et al., 2024b; Zou et al., 2023): even when the downstream dataset contains only a few harmful samples or consists entirely of seemingly safe ones, fine-tuning can still erode the LLM’s built-in safety mechanisms, causing the fine-tuned model to generate harmful content, which raises a fundamental question: *how can we enable safe and personalized fine-tuning without sacrificing alignment guarantees?*

To this end, various strategies have been proposed, ranging from alignment-stage defense (Wang et al., 2025a; Zheng et al., 2025; Huang, 2025), fine-tuning-stage defense (Wang et al., 2024; Lyu et al., 2024; Huang et al., 2024b; Qi et al., 2024a) to post-processing-stage defense (Wang et al., 2025b; Gong et al., 2025; Yang et al., 2025; Lu et al., 2025), where data-centric approaches that purify the fine-tuning stage are particularly promising. These approaches span a wide spectrum, including heuristic filtering (Xie et al., 2023a; Choi et al., 2024), bi-level reweighting (Shen et al., 2025), and adaptive probabilistic scheduling (Hu et al., 2025). However, these methods face two fundamental limitations. First, they predominantly operate at an instance level, assessing the utility of individual data points in isolation while neglecting the global geometry of the data manifold. Second, they typically adopt a single-reference paradigm: they aim to align the model with safe examples (attraction) but fail to explicitly utilize harmful samples as a negative reference for active repulsion. Consequently, lacking these structural constraints, even advanced reweighting schemes may struggle to robustly steer the model’s underlying distribution away from harmful regions.

To bridge this critical gap, we introduce Safety Optimal Transport (SOT), a novel framework for safeguarding LLMs during fine-tuning. At its core, SOT is a principled weight-learning framework that

\* These authors contributed equally.

† Corresponding author.

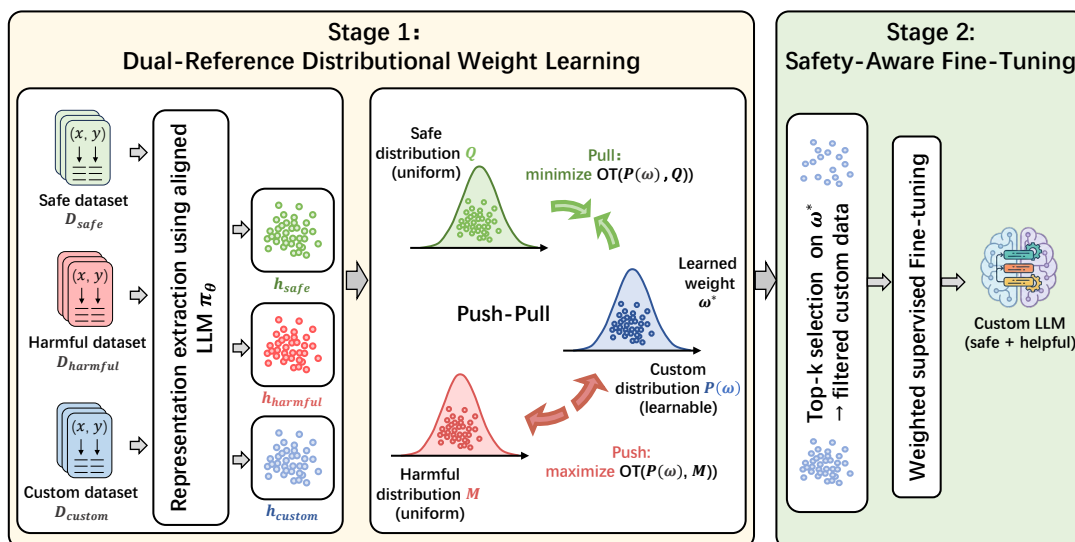


Figure 1: Overview of the SOT framework. Stage 1 (Dual-Reference Distributional Weight Learning) utilizes a dual-reference “push-pull” Optimal Transport strategy to learn sample weights by aligning with safe anchors and repelling harmful ones. Stage 2 (Safety-Aware Fine-Tuning) applies these weights for Top-k selection and weighted SFT to produce a final safe and helpful LLM.

learns an importance weight for each sample in the user-provided dataset from the perspective of distributional alignment, beyond the existing instance-level estimation. As shown in Fig. 1, specifically, we can represent the downstream task dataset as a discrete empirical distribution over all samples within it, where we view the to-be-learned weight vector as its probability measure. Besides, we can represent a safe dataset as a discrete empirical safe distribution over all samples within it, and also represent a popular harmful dataset as a discrete empirical harmful distribution over all samples within it. To optimize the weights, we define a clear objective based on a “push-pull” dynamic: the learned weights must “pull” the downstream data distribution closer to a trusted safe reference, while simultaneously “pushing” it away from a harmful one. Therefore, the learning of a weight vector in the downstream task dataset can be formulated as the process of learning the distribution to be as close to the safe distribution as possible, and to be as far away from the harmful distribution, which can be facilitated with the OT distance (Cuturi, 2013). Notably, the cost function plays a paramount role when learning the transport plan for OT, where we use the safety-aligned language model’s last-layer representations of samples to design it. That is to say, we can disengage the dependence of the weight learning on the to-be-fine-tuned LLMs. This “push-pull” mechanism ensures a robust weight assignment that purifies the training data. Given the optimized weight vector, we can select the samples

with large weights and use them to perform the downstream task fine-tuning. Our main contributions are:

1. We propose a novel weight-learning methodology based on Optimal Transport, which reframes the safety problem from an instance-level filtering challenge to a more holistic, distribution-level alignment task.
2. Our weight-learning is guided by a unique dual-reference mechanism that not only “pulls” the data distribution towards a safe ideal but also actively “pushes” it away from a harmful one.
3. Extensive experiments demonstrate that fine-tuning with our SOT-learned weights significantly enhances model safety without compromising performance on downstream tasks.

## 2 Related Work

**LLM Safety Alignment.** The LLM safety alignment aims to ensure generations towards helpfulness and harmlessness, where early frameworks established Reinforcement Learning from Human Feedback (RLHF) based on reward models (Li et al., 2025g,d; Duan et al., 2026) as foundational: (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022). Subsequent work expanded RLHF to Constitutional AI (Coan and Surden, 2025) and Direct Preference Optimization (DPO) (Rafailov et al., 2023). However, adversarial testing reveals persistent vulnerabilities. Human red teaming uncovered significantly more harmful outputs than automated methods (Ganguli et al., 2022; Du

et al., 2025b; Li et al., 2025e), while gradient-based attacks (Perez et al., 2022) and LLM-powered adversaries (Mazeika, 2024) exposed new risks. To address these, safety benchmarking efforts standardized evaluations (e.g., SafetyBench (Zhang et al., 2023) and DecodingTrust (Wang et al., 2023)), and corrective techniques emerged, including activation steering (Subramaniam et al., 2025) and unlearning methods (Jang et al., 2022).

**Harmful Fine-tuning.** Safety-aligned LLMs remain vulnerable during fine-tuning, where implicit or explicit harmful samples can compromise model alignment (Zhan et al., 2024; Lermen and Rogers-Smith, 2023; Yi et al., 2024; Yang et al., 2023a; Zhao et al., 2025). Defensive strategies against this vulnerability generally span three stages: (i) Alignment-stage defense, reinforcing the base LLM’s safety prior to training (Wang et al., 2025a; Zheng et al., 2025; Huang, 2025; Tamirisa et al., 2024; Huang et al., 2024c; Chen et al., 2025b; Rosati et al., 2024; Chen et al., 2025a; Dionicio et al., 2025; Perin et al., 2025); (ii) Fine-tuning-stage defense, mitigating risks during task adaptation (Wang et al., 2024; Lyu et al., 2024; Huang et al., 2024b; Qi et al., 2024a; Eiras et al., 2024; Zhang et al., 2024; Li et al., 2025c,b; Kim et al., 2025; Bianchi et al., 2024; Huang et al., 2024b; Choi et al., 2024; Li et al., 2026); (iii) Post-processing-stage defense, purifying or “unlearning” harmful behaviors after training (Wang et al., 2025b; Gong et al., 2025; Yang et al., 2025; Lu et al., 2025; Zhou et al., 2025, 2024; Huang et al., 2024a; Djuhera et al., 2025).

Our work belongs to the second group, which aims to solve the problem from the view of data-centric selection. While recent advancements that range from heuristic filtering (Choi et al., 2024; Xie et al., 2023a) to model-loop bilevel optimization (Shen et al., 2025) and Bayesian inference (Hu et al., 2025) have improved selection efficacy, they predominantly share a single-reference, instance-level paradigm. Specifically, these methods primarily focus on estimating the safety likelihood of each sample individually to “attract” the model toward a safe prior, which neglects to explicitly leverage harmful distributions as a negative reference for active repulsion. By introducing a dual-reference “push-pull” mechanism, SOT not only gravitates toward safe distributions but also explicitly repels harmful ones, establishing a geometric safety boundary that is more robust than methods

relying solely on positive selection.

**Sample Selection for LLMs.** Sample selection aims to curate high-quality subsets to improve model performance and safety. Current approaches primarily rely on LLM-based grading (Chen et al., 2024; Lu et al., 2023; Li et al., 2025f) or statistical/influence-based heuristics (Cao et al., 2023; Xie et al., 2023a; Xia et al., 2024) to identify “generally good” data. However, these methods often focus on individual sample quality (e.g., diversity or complexity) while neglecting the global distribution mismatch between custom fine-tuning data and the original alignment distribution. In contrast, we formulate sample selection as an Optimal Transport problem, explicitly aligning the custom data distribution with reference distributions. Our approach not only provides interpretable transport weights for targeted filtering without external dependencies but also ensures a more robust safety boundary by minimizing distributional erosion.

### 3 Preliminary

**Supervised Fine-Tuning for LLM.** Let  $\pi_\theta$  denote a pre-trained LLM parameterized by  $\theta$ . Given a downstream dataset as  $\mathcal{D}_{\text{custom}} = \{(\mathbf{x}_i, \mathbf{y}_i)_{\text{custom}}\}_{i=1}^{N_c}$  with  $N_c$  training pairs, each  $(\mathbf{x}, \mathbf{y})$  includes an input sequence  $\mathbf{x}$  (e.g., an instruction or question) and a target response  $\mathbf{y}$ . SFT mainly aims to adapt  $\pi_\theta$  to the specific tasks by minimizing the Negative Log-Likelihood (NLL) loss over  $\mathcal{D}_{\text{custom}}$ :

$$\begin{aligned} \mathcal{L}_{\text{SFT}}(\theta) &= -\mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}_{\text{custom}}} [\log \pi_\theta(\mathbf{y} | \mathbf{x})] \\ &= -\frac{1}{N_c} \sum_{i=1}^{N_c} \frac{1}{|\mathbf{y}_i|} \sum_{j=1}^{|\mathbf{y}_i|} \log \pi_\theta(y_{i,j} | \mathbf{x}_i, \mathbf{y}_{i, < j}). \end{aligned} \quad (1)$$

By optimizing Eq. 1, the model  $\pi_\theta$  mimics the response style, format, and knowledge in the downstream data, thereby becoming a specialized model tailored for the target application. However, the efficiency of SFT is critically dependent on the quality of the training data (Perez et al., 2022; Huang et al., 2024a; Du et al., 2025a). If  $\mathcal{D}_{\text{custom}}$  is contaminated with unsafe or low-quality samples,  $\pi_\theta$  will inadvertently learn to replicate these harmful behaviors, severely compromising the intended safety alignment.

**Optimal Transport.** OT measures the minimal transport cost of transforming one probability distribution into another, where we mainly consider the commonly-used discrete form in this

manuscript (Guo et al., 2022; Gao et al., 2023; Li et al., 2025h). Denote two discrete probability distributions  $P = \sum_{i=1}^n a_i \delta_{\mathbf{x}_i}$  and  $Q = \sum_{j=1}^m b_j \delta_{\mathbf{x}_j}$ , where both  $\mathbf{a}$  and  $\mathbf{b}$  are discrete probability vectors summing to 1,  $\mathbf{x}_i$  and  $\mathbf{x}_j$  are the supports of the two distributions respectively, and  $\delta$  is a Dirac function. The OT distance is formulated as follows:

$$\text{OT}(P, Q) = \min_{\mathbf{T} \in \Pi(P, Q)} \langle \mathbf{T}, \mathbf{C} \rangle, \quad (2)$$

where  $\mathbf{C} \in \mathbb{R}_{\geq 0}^{n \times m}$  is the cost matrix and each  $C_{ij}$  measures the distance between  $\mathbf{x}_i$  and  $\mathbf{x}_j$ . The transport plan  $\mathbf{T} \in \mathbb{R}_{\geq 0}^{n \times m}$  satisfies  $\Pi(P, Q) := \left\{ \mathbf{T} \mid \sum_{i=1}^n T_{ij} = b_j, \sum_{j=1}^m T_{ij} = a_i \right\}$ . To reduce the computational cost in the above optimization problem, the Sinkhorn algorithm (Cuturi, 2013) introduces a popular entropic regularization term  $H = -\sum_{ij} T_{ij} \ln T_{ij}$ .

## 4 Methodology

We introduce Safety Optimal Transport (SOT), which reformulates safe fine-tuning as a distribution level alignment task. As shown in Fig. 1, SOT includes two stages: dual-reference distributional weight learning and safety-aware fine-tuning.

### 4.1 Dual-Reference Distributional Weight Learning

The core objective of SOT is to learn a global weight vector  $\mathbf{w}$  that assigns importance to each instance in the downstream dataset  $\mathcal{D}_{\text{custom}} = \{(x_i, y_i)\}_{i=1}^{N_c}$ . Rather than treating data selection as a collection of independent binary decisions that often fails to capture collective semantic relationships, we reformulate the problem as learning a continuous probability measure. Specifically, we represent the weighted dataset as a discrete empirical distribution

$$P(\mathbf{w}) = \sum_{i=1}^{N_c} w_i \delta_{(\mathbf{x}_i, \mathbf{y}_i)}, \quad (3)$$

where weight vector  $\mathbf{w} \in \Delta^{N_c-1}$  satisfies  $\sum w_i = 1$  and  $w_i \geq 0$ , and  $\delta_{(\mathbf{x}_i, \mathbf{y}_i)}$  denotes the Dirac mass at location  $(\mathbf{x}_i, \mathbf{y}_i)$ . Optimizing  $P(\mathbf{w})$  requires explicit geometric guidance, we design a dual-reference strategy. (1) Task-Specific Alignment (Pull): The distribution  $P(\mathbf{w})$  must align with trusted data from the target domain to preserve downstream utility while ensuring safety. We construct a distribution  $Q$  using a small, verified subset  $\mathcal{D}_{\text{safe}}$  (e.g., a few dozen samples) sampled

from the downstream task distribution, which acts as a high-precision anchor to ensure the learned weights prioritize samples that are both safe and utility-preserving. (2) General Harm Repulsion (Push): The distribution  $P(\mathbf{w})$  must actively maximize its distance from broad toxicity patterns to establish a robust geometric safety margin. We employ a subset of a publicly available harmful dataset  $\mathcal{D}_{\text{harmful}}$  (e.g., red-teaming benchmarks) to represent generic toxicity patterns and build the distribution  $M$ , which serves as a repulsive source without requiring domain-specific harmful data collection. Both references are formulated as uniform distributions:

$$Q = \sum_{j=1}^{N_s} \frac{1}{N_s} \delta_{(\mathbf{x}_j, \mathbf{y}_j)^s}, M = \sum_{k=1}^{N_h} \frac{1}{N_h} \delta_{(\mathbf{x}_k, \mathbf{y}_k)^h}, \quad (4)$$

where each  $(\mathbf{x}_j, \mathbf{y}_j)^s \sim \mathcal{D}_{\text{safe}}$  and  $(\mathbf{x}_j, \mathbf{y}_j)^h \sim \mathcal{D}_{\text{harmful}}$ .  $N_s$  and  $N_h$  denote the number of samples in each corresponding dataset.

**The Unified Push-Pull Objective.** To achieve both robust safety and high utility, SOT integrates these two forces into a unified objective. We optimize  $\mathbf{w}$  to “pull” the distribution towards safety while simultaneously “pushing” it away from harm. We have the final objective formulated as:

$$\min_{\mathbf{w}} \left[ (1-\lambda) \text{OT}(P(\mathbf{w}), Q) - \lambda \text{OT}(P(\mathbf{w}), M) \right], \quad (5)$$

where  $\lambda \in [0, 1]$  controls the trade-off between safety alignment and harmful repulsion. By incorporating the OT distance, Eq. 5 reframes weight learning as a geometric transport problem, which evaluates a sample’s quality not by its isolated attributes, but by its transport cost relative to the entire safe manifold ( $Q$ ) and harmful manifold ( $M$ ). The Push force ( $\text{OT}(P(\mathbf{w}), M)$ ) acts as a discriminator, compressing the weights of samples that are semantically close to harmful samples (including camouflaged hard negatives) towards zero. The Pull force  $\text{OT}(P(\mathbf{w}), Q)$  acts as a relevance ranker, ensuring that the remaining safe samples are weighted according to their semantic alignment with the task-specific anchor. While our “push-pull” objective conceptually resembles contrastive learning, SOT fundamentally differs in two ways. First, instead of relying on explicitly constructed instance pairs, SOT aligns entire empirical distributions holistically via Optimal Transport. Second, rather than directly modifying the underlying representation geometry, SOT focuses strictly on weight

learning. It leverages the transport plan to induce sample weights, actively purifying the fine-tuning data without distorting the pre-trained model’s intrinsic semantic space.

To make Eq. 5 computationally tractable and differentiable, we adopt the Entropy-regularized Optimal Transport (Cuturi, 2013). We thus optimize the following objective:

$$\begin{aligned} \min_{\mathbf{w}} \mathcal{L}_{\text{SOT}} &= (1 - \lambda) \langle \mathbf{T}^{s,*}, \mathbf{C}^s \rangle - \lambda \langle \mathbf{T}^{h,*}, \mathbf{C}^h \rangle, \\ \text{s.t. } \mathbf{T}^{s,*} &= \arg \min_{\mathbf{T}^s \in \Pi(P(\mathbf{w}), Q)} \\ &\quad \langle \mathbf{T}^s, \mathbf{C}^s \rangle - \epsilon H(\mathbf{T}^s), \\ \mathbf{T}^{h,*} &= \arg \min_{\mathbf{T}^h \in \Pi(P(\mathbf{w}), M)} \\ &\quad \langle \mathbf{T}^h, \mathbf{C}^h \rangle - \epsilon H(\mathbf{T}^h), \end{aligned} \quad (6)$$

where  $\epsilon > 0$  controls the weight of the entropic regularization and  $H(\cdot)$  computes the entropy. The cost matrix (e.g.,  $\mathbf{C}^s \in \mathbb{R}_{\geq 0}^{N_c \times N_s}$  between  $P(\mathbf{w})$  and  $Q$ ) quantifies the semantic discrepancy using the Cosine Distance:  $C_{ij}^s = 1 - \frac{\mathbf{z}_i^\top \mathbf{z}_j}{\|\mathbf{z}_i\|_2 \|\mathbf{z}_j\|_2}$ . For each sample pair, we construct the input context [Human:  $x_i$ ; Assistant:  $y_i$ ] and extract the representation  $\mathbf{z}_i$  from the frozen safety-aligned model  $\pi_\theta$  before the fine-tuning process. The transport plan (e.g.,  $\mathbf{T}^s \in \mathbb{R}_{\geq 0}^{N_c \times N_s}$ ) should satisfy  $\Pi(P(\mathbf{w}), Q) := \left\{ \mathbf{T}^s \mid \sum_{i=1}^{N_c} T_{ij}^s = 1/N_s, \sum_{j=1}^{N_s} T_{ij}^s = w_i \right\}$ . To optimize  $\mathbf{w}$  at each step, we first compute the optimal transport plans  $\mathbf{T}^{s,*}$  and  $\mathbf{T}^{h,*}$  by optimizing the corresponding entropy regularized OT loss, respectively. Since the optimal transport plans are the functions of  $\mathbf{w}$ , they are then utilized to compute the objective function  $\mathcal{L}_{\text{SOT}}$ . Finally, we update  $\mathbf{w}$  (parameterized via softmax) by minimizing the global Push-Pull objective  $\mathcal{L}_{\text{SOT}}$  via automatic differentiation using gradient descent.

## 4.2 Safety-Aware Fine-Tuning

**Hard Denoising via Top-K Selection.** Given the optimized weight vector  $\mathbf{w}^*$  obtained by minimizing the final objective in Eq. 6, low-quality samples in  $\mathcal{D}_{\text{custom}}$  are naturally assigned lower weights, while higher weights correspond to more helpful and safer examples. Considering non-zero gradients from severe harmful samples might accumulate to degrade safety alignment, we apply a hard truncation strategy to filter out potentially toxic tails and improve training efficiency. We select the subset  $\mathcal{D}'_{\text{custom}}$  corresponding to the Top- $K$  indices

---

### Algorithm 1 SOT Alignment Framework

---

```

1: Require: Datasets  $\{\mathcal{D}_{\text{custom}}, \mathcal{D}_{\text{safe}}, \mathcal{D}_{\text{harmful}}\}$ , LLM  $\pi_\theta$ , hyper-parameters.
2: Extract representations of  $(\mathbf{x}, \mathbf{y})$  within  $\{\mathcal{D}_{\text{custom}}, \mathcal{D}_{\text{safe}}, \mathcal{D}_{\text{harmful}}\}$  from  $\pi_\theta$  by concatenating  $\mathbf{x}$  and  $\mathbf{y}$  into one input context.
3: Build  $\{P(\mathbf{w}), Q, M\}$  according to Eq. 3 with uniformly initialized weight  $\mathbf{w}$ , and Eq. 4, respectively.
4: for  $e_1$  in  $[1, E_1]$  do
5:   Optimize  $\mathbf{T}^{s,*}$  and  $\mathbf{T}^{h,*}$  according to Eq. 6.
6:   Optimize  $\mathbf{w}$  by minimizing  $\mathcal{L}_{\text{SOT}}$  according to Eq. 6.
7: end for
8: Build  $\mathcal{D}'_{\text{custom}}$  by performing Top- $K$  selection on ranked  $\mathbf{w}$  in Eq. 7.
9: for  $e_2$  in  $[1, E_2]$  do
10:   Update  $\pi_\theta$  based on  $\mathcal{D}'_{\text{custom}}$  by minimizing  $\mathcal{L}_{\text{custom}}(\theta)$  in Eq. 8.
11: end for

```

---

of  $\mathbf{w}^*$  by:

$$\mathcal{D}'_{\text{custom}} = \left\{ (\mathbf{x}_i, \mathbf{y}_i, w'_i)^{\text{custom}} \mid i \in \text{Top-}K(\mathbf{w}^*) \right\}, \quad (7)$$

where  $K$  controls the size of the selected dataset,  $\text{Top-}K(\mathbf{w}^*)$  denotes the set of indices corresponding to the  $K$  largest values in the learned weight vector  $\mathbf{w}^*$ . Within the selected safe subset, samples still vary in utility and relevance. We re-normalize the weights:  $w'_i = \frac{\exp(w_i)}{\sum_{j \in \text{Top-}K(\mathbf{w}^*)} \exp(w_j)}$ .

### Soft Emphasis via Weighted LLM Finetune.

After constructing the Top- $K$  subset  $\mathcal{D}'_{\text{custom}}$  and its corresponding weight vector  $\mathbf{w}'$ , we fine-tune the target LLM  $\pi_\theta$  by reweighting the standard supervised fine-tuning loss. Formally, the final training loss is defined as:

$$\mathcal{L}_{\text{custom}}(\theta) = - \sum_{i=1}^K w'_i \sum_{j=1}^{|\mathbf{y}_i|} \frac{1}{|\mathbf{y}_i|} \log \pi_\theta(y_{ij} \mid \mathbf{x}_i, y_{i,<j}), \quad (8)$$

where the re-normalized weight  $w'_i$  acts as a safety coefficient, adaptively modulating the gradient magnitude for each sample, which ensures that the model updates are predominantly driven by high-utility, safety-aligned instances. Following Shen et al. (2025), we employ LoRA (Hu et al., 2021) for efficient fine-tuning by injecting low-rank matrices while freezing the backbone, which effectively mitigates overfitting and integrates the safety-aligned sample weights. In summary, our framework introduces a systematic approach to data purification for safer SFT. By minimizing Optimal Transport distance in Eq. 6 to optimize a sample-wise weight vector  $\mathbf{w}$  guided by reference distributions, we effectively redistribute the downstream dataset  $\mathcal{D}_{\text{custom}}$  towards safety and helpfulness. As a benefit, not only does the learned vector provide a reliable criterion for filtering out low-quality data, but more importantly, it facilitates a weighted

Table 1: Experimental results (HIS, HpS, Avg.) on Meta-Llama-3.1-8B-Instruct (top) and Qwen3-8B (down), respectively. We highlight the best result in **Red**.

	Dataset Method	SLIMORCA			AGNews			MetaMathQA			GSM8K		
		HIS↑	HpS↑	Avg.↑	HIS↑	Acc↑	Avg.↑	HIS↑	HpS↑	Avg.↑	HIS↑	HpS↑	Avg.↑
Meta-Llama3.1-8B	Standard SFT	0.574	3.764	0.663	0.709	<b>0.876</b>	0.793	0.629	3.457	0.660	0.597	3.954	0.694
	Salora (Li et al., 2025b)	0.575	3.781	0.666	0.690	0.875	0.783	0.624	3.421	0.654	0.588	3.997	0.694
	SafeInstr (Bianchi et al., 2024)	0.584	3.767	0.669	0.744	0.872	0.808	0.680	3.676	0.708	0.632	4.060	0.722
	Random Selection	0.640	3.753	0.695	0.761	0.864	0.813	0.601	3.291	0.630	0.518	4.102	0.669
	ALPAGASUS (Chen et al., 2024)	0.648	3.769	0.701	0.690	0.856	0.773	0.724	<b>3.884</b>	0.750	0.925	4.103	0.873
	DSIR (Xie et al., 2023a)	0.765	3.806	0.763	0.910	0.757	0.833	0.925	3.485	0.811	0.924	4.109	0.873
	SAFT (Choi et al., 2024)	0.605	3.802	0.682	0.639	0.868	0.754	0.665	3.612	0.694	0.618	4.055	0.715
	BDS (Hu et al., 2025)	0.352	3.621	0.538	0.749	0.821	0.785	0.751	3.559	0.731	0.868	4.015	0.835
	SEAL (Shen et al., 2025)	0.639	3.792	0.699	0.712	0.871	0.792	0.705	3.479	0.700	0.608	4.112	0.715
	<b>SOT</b>	<b>0.803</b>	<b>3.895</b>	<b>0.791</b>	<b>0.946</b>	0.859	<b>0.903</b>	<b>0.967</b>	3.602	<b>0.844</b>	<b>0.926</b>	<b>4.114</b>	<b>0.874</b>
Qwen3-8B	Standard SFT	0.351	<b>3.633</b>	0.539	0.374	<b>0.904</b>	0.639	0.330	3.614	0.526	0.335	3.890	0.557
	Salora (Li et al., 2025b)	0.362	3.607	0.542	0.368	<b>0.904</b>	0.636	0.322	3.611	0.522	0.340	3.877	0.558
	SafeInstr (Bianchi et al., 2024)	0.388	3.624	0.556	0.363	0.901	0.632	0.330	3.619	0.527	0.316	4.004	0.559
	Random Selection	0.365	3.544	0.537	0.356	0.902	0.629	0.329	3.456	0.510	0.335	3.912	0.559
	ALPAGASUS (Chen et al., 2024)	0.768	3.396	0.723	0.377	0.897	0.637	0.338	<b>3.678</b>	0.537	0.722	3.992	0.760
	DSIR (Xie et al., 2023a)	0.832	3.475	0.763	0.700	0.827	0.764	0.719	3.017	0.661	0.758	3.115	0.690
	SAFT (Choi et al., 2024)	0.344	3.589	0.531	0.372	0.891	0.632	0.347	3.668	0.540	0.328	3.942	0.558
	BDS (Hu et al., 2025)	0.609	3.611	0.666	0.630	0.890	0.760	0.370	3.771	0.562	0.438	4.043	0.623
	SEAL (Shen et al., 2025)	0.349	3.630	0.538	0.341	0.895	0.618	0.701	3.018	0.652	0.346	3.941	0.567
	<b>SOT</b>	<b>0.834</b>	3.559	<b>0.773</b>	<b>0.761</b>	0.893	<b>0.827</b>	<b>0.787</b>	3.472	<b>0.741</b>	<b>0.769</b>	<b>4.122</b>	<b>0.796</b>

fine-tuning process that compels the model to prioritize high-quality, safe examples. Consequently, SOT achieves better alignment in both helpfulness and safety by strategically amplifying the learning signal from valuable data while mitigating the influence of harmful instances. We highlight our training process in Algorithm 1.

## 5 Experiments

### 5.1 Setup

**Datasets.** Following Shen et al. (2025); Huang et al. (2024b), we use SLIMORCA (Lian et al., 2023), MetaMathQA (Yu et al., 2023), AGNews (Zhang et al., 2015), GSM8K (Cobbe et al., 2021) and the specialized legal dataset LegalBenchQA (Menta, 2023), which is derived from the original LegalBench benchmark (Guha et al., 2023) as fine-tuning tasks. We use the red-team-attempts subset of HH dataset (Bai et al., 2022) as the source of harmful data in the fine-tuning dataset. Within a fine-tuning dataset containing a total of  $n$  samples, we set the mixture of safe / harmful data as  $1 - p : p$ , where the default setting is  $p = 0.1$  and  $n = 5000$ . For the harmful dataset, we employ a subset of the first 5,000 harmful samples from BeaverTails (Ji et al., 2023), while the safe dataset comprises 50 instances randomly sampled from the corresponding custom downstream dataset.

**Experimental setting.** For SOT, we set  $\lambda$  as 0.5 and fix  $\epsilon = 0.1$  in computing the transport plan. We extract the last-token hidden state in the last layer from the frozen safety-aligned model  $\pi_\theta$  as

the representation  $\mathbf{z}$ . For a fair comparison, we select 80% data (i.e.,  $K = 4000$ ) as the fine-tuning data following Shen et al. (2025). We adopt LoRA for accelerating the LLM fine-tuning and train  $\pi_\theta$  for 2 epochs. Detailed hyper-parameters and setups can be found in Appendix A.1 and Appendix A.2.

**Baselines.** To fairly prove the effectiveness of our method, we consider the following baselines: 1) Standard SFT. 2) Data Selection Methods, including random selection, ALPAGASUS (Chen et al., 2024) and DSIR (Xie et al., 2023a). 3) Safety Fine-tuning Methods based on Data Selection and Reweighting, including SAFT (Choi et al., 2024), BDS (Hu et al., 2025) and SEAL (Shen et al., 2025). 4) Structure-Aware Methods and Instruction Fine-Tuning Methods, including Salora (Li et al., 2025b) and SafeInstr (Bianchi et al., 2024). More details of these baselines are in Appendix A.3.

**Models and Evaluation.** Following Shen et al. (2025); Li et al. (2025a), we evaluate the effectiveness of SOT on Meta-Llama-3.1-8B-Instruct (Dubey et al., 2024), Meta-Llama-3-8B-Instruct (Dubey et al., 2024), Qwen3-8B (Team, 2025b), and gemma-3-4b-it (Team, 2025a). We construct a Helpfulness test set and a Harmfulness test set to evaluate the fine-tuned LLM, both including 1000 samples from the corresponding downstream task datasets and red-team-attempts dataset. We define Helpfulness Score (HpS) and Harmlessness Score (HIS) to assess the performance and safety of fine-tuned LLM by adopting a popular LLM-as-a-judgment evaluation strategy. For the

AGNews dataset, we use Accuracy (Acc.) to measure the model’s classification performance. Finally, we use Average Score (Avg.) to measure the overall performance of the model in terms of safety and usefulness. More details can be found in Appendix A.4.

## 5.2 Main Results

**Experimental Results.** Tab. 1 shows that SOT consistently achieves significant performance improvement across both Meta-Llama-3.1-8B-Instruct and Qwen-3-8B families, demonstrating robust generalization capabilities. Specifically, SOT yields the highest Avg. across all evaluated datasets, including SLIMORCA, AGNews, MetaMathQA, and GSM8K. Critically, SOT attains the lowest HIS in every scenario, significantly outperforming competitive baselines like DSIR and SEAL in safety alignment. While standard SFT or ALPAGASUS occasionally achieve high utility scores (e.g., accuracy on AGNews or HpS on MetaMathQA), they often do so at a substantial cost to safety. In contrast, SOT maintains highly competitive and often superior HpS, effectively navigating the safety-utility tradeoff. This universal improvement across diverse downstream domains and distinct model architectures underscores SOT’s efficacy as a balanced and scalable solution for real-world LLM usage. In addition, we provide results on Llama3-8B-Instruct and Gemma3-4B-it in Appendix A.5, where we find the consistently improved performance compared with strong baselines.

### Specialized Domain Generalization on Law.

Beyond general benchmarks, SOT demonstrates exceptional efficacy in specialized domains, as evidenced by results on the LegalBench-QA dataset. As shown in Tab. 2, SOT achieves a better safety-utility trade-off, securing the highest average score (0.834) while maintaining a significantly higher Harmlessness Score (0.895) compared to all tuned baselines. Notably, while competitive methods like DSIR may achieve higher Helpfulness Scores, they often incur much higher safety risks. Furthermore, to isolate the effectiveness of our optimization mechanism from the choice of reference data, we evaluate a variant of SOT guided by the third-party safety dataset from BDS. SOT’s ability to substantially improve safety with minimal impact on utility in a high-stakes legal context highlights its practical robustness and adaptability across both general and specific domains.

Table 2: Experimental results of Meta-Llama-3.1-8B-Instruct on the LegalBench-QA dataset, best results in Red.

Method	HIS↑	HpS↑	Avg.↑
Standard SFT	0.741	3.810	0.752
Salora (Li et al., 2025b)	0.732	3.798	0.746
SafeInstr (Bianchi et al., 2024)	0.715	3.832	0.741
Random Selection	0.712	3.838	0.740
ALPAGASUS (Chen et al., 2024)	0.695	3.782	0.726
DSIR (Xie et al., 2023a)	0.825	<b>3.935</b>	0.806
SAFT (Choi et al., 2024)	0.700	3.817	0.732
BDS (Hu et al., 2025)	0.849	3.794	0.804
SEAL (Shen et al., 2025)	0.693	3.836	0.730
SOT (w/ BDS safety (Hu et al., 2025))	<b>0.901</b>	3.704	0.821
<b>SOT</b>	0.895	3.865	<b>0.834</b>

## 5.3 Additional Analysis

### Effectiveness of the proposed modules.

Tab. 3 presents a holistic ablation study examining the individual contributions of SOT’s components—specifically its objectives and fine-tuning strategies—alongside the reference data sources.

(1) First and foremost, the results demonstrate that the “Pull” and “Push” objectives are mutually indispensable. While the “Pull Only” configuration (HIS: 0.776) establishes a baseline for safety convergence, and “Push Only” (HIS: 0.559) struggles without a constructive target, the superior safety-utility trade-off is achieved only when both forces are combined (HIS: 0.803). This confirms that neither mechanism is sufficient in isolation: the “Pull” force provides the necessary guidance towards the safe manifold, while the “Push” force serves as an essential negative constraint to actively purify adversarial spaces that a positive reference alone cannot address. Conceptually, our dual-reference objective is not an empirical heuristic, but a principled mechanism that enforces a geometric margin in the representation space. By optimizing the Eq. 5 SOT induces a structural separation. The mechanism enforces a contraction toward the safe distribution ( $Q$ ) and an expansion away from the harmful distribution ( $M$ ). Specifically, the “push” term actively increases the Wasserstein distance to the harmful supports. Under entropy-regularized Optimal Transport, the transport plan remains smooth and stable, allowing this expansion to implicitly enlarge the safety margin boundary. Therefore, SOT provides a robust distribution-level separation.

(2) Beyond the learning objective, our analysis of the fine-tuning strategy reveals a strong synergy between filtering and reweighting. While the “Top-K Only” approach effectively acts as a hard denoiser to truncate toxic long-tail samples (HIS: 0.205), it is the integration with “Reweighting”—which

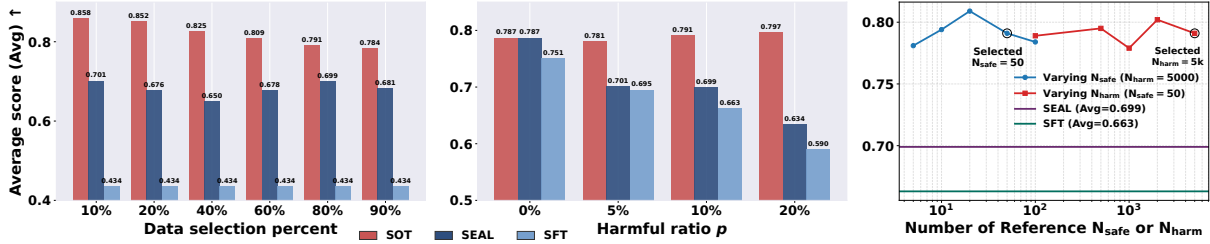


Figure 2: The left figure shows results for different data selection rates, the middle figure presents results for fine-tuned datasets containing varying proportions of harmful samples, the right figure shows the reference data efficiency of SOT using Meta-Llama-3.1-8B-Instruct on the SLIMORCA dataset.

Table 3: Ablation study on method and data resources. ‘-’ indicates the module is disabled. In the *Objective* columns, ‘✓’ represents the default configuration (safe data: task-specific data, harm data: BeaverTails). Variants use subscripts to denote alternative sources: ✓<sub>Gen</sub> = General Safe (BeaverTails), ✓<sub>Adv</sub> = Adversarial Harm from Sheshadri et al. (2024).

Configuration	Objective		Strategy		Metrics		
	Pull	Push	Top-K	Reweight	HIS ↑	HpS ↑	Avg. ↑
Standard SFT	-	-	-	-	0.574	3.764	0.663
SEAL	-	-	✓	-	0.639	3.792	0.699
<i>Method Ablation</i>							
Pull Only	✓	-	✓	✓	0.776	3.780	0.766
Push Only	-	✓	✓	✓	0.559	3.779	0.657
Reweight Only	✓	✓	-	✓	0.623	3.745	0.686
Top-K Only	✓	✓	✓	-	0.795	3.834	0.781
<i>Data Source Ablation (SOT Framework)</i>							
w/ General Safe	✓ <sub>Gen</sub>	✓	✓	✓	0.457	3.816	0.610
w/ Another Harm	✓	✓ <sub>Adv</sub>	✓	✓	0.806	3.836	0.787
<b>SOT (Ours)</b>	✓	✓	✓	✓	<b>0.803</b>	<b>3.895</b>	<b>0.791</b>

offers a smoother optimization landscape—that allows the full SOT strategy to maximize the Average Score. This combined approach ensures that while the most harmful instances are removed, the remaining safe subset is further refined by up-weighting high-quality samples.

(3) Finally, the study highlights the distinct nature of our distribution matching dependencies. We find that for the ‘‘Pull’’ mechanism, replacing the task-specific safe anchor with a general safe dataset (e.g., BeaverTails) leads to a sharp degradation in safety. In contrast, the ‘‘Push’’ mechanism demonstrates remarkable robustness: replacing the default harmful dataset with another adversarial set yields nearly identical performance (Avg 0.787 vs 0.791), demonstrating robustness to domain variation. This indicates that while the guidance towards safety must be precise and task-aligned, the repulsion from harmfulness generalizes effectively across different manifestations of toxicity. We give more granular analysis in Appendix A.6.

**Impact of Selection and Harmful Ratios.** As illustrated in Fig. 2 (Left), SOT maintains a robust Average Score (Avg.) of approximately 0.8 across selection rates ranging from 10% to 90%, consistently surpassing SEAL and SFT. Notably,

the performance gap widens as the selection rate decreases; this validates that SOT excels at prioritizing the highest-quality samples, ensuring superior alignment even with a minimal data budget. The framework’s resilience to data contamination is further evidenced in Fig. 2 (Middle). While baselines suffer precipitous degradation as the harmful ratio  $p$  increases, SOT sustains a stable profile near 0.8, even under severe contamination. Crucially, SOT outperforms standard SFT even in pristine regimes ( $p = 0$ ), highlighting its ability to amplify high-utility samples through safety-aware reweighting independent of noise removal. This stability across the spectrum demonstrates SOT’s dual capacity for precise data identification and robust learning in adversarial scenarios.

**Data Efficiency and Asymmetry.** Fig. 2 (Right) reveals a distinct asymmetry in reference data requirements. The rapid saturation of the blue curve demonstrates that the safe anchor provides directional guidance, achieving near-optimal performance with as few as  $N_{\text{safe}} \in [5, 50]$  samples. Notably, even with 5 safe samples, SOT substantially outperforms baselines. In contrast, rather than exhibiting strict scale dependence, the ‘‘Push’’ mechanism (red curve) relies on a sufficiently diverse harmful reference to stabilize the adversarial boundary. This reflects the necessity of adequately covering the high-variance manifold of toxicity to provide reliable directional repulsion. Despite these differing demands, SOT consistently outperforms baselines (SEAL and SFT) across all settings. This confirms that SOT efficiently anchors safety with minimal clean demonstrations while leveraging broad negative coverage to fortify the adversarial boundary.

## 5.4 Conceptual and Empirical Justification for Optimal Transport

Conceptually, the primary objective of Equation (5) is to reshape the downstream empirical distri-

bution relative to safe ( $Q$ ) and harmful ( $M$ ) reference manifolds. In this context, Optimal Transport (OT) offers critical advantages over standard scalar divergence measures, such as KL-divergence or f-divergence, primarily through its geometry awareness and explicit correspondence.

Traditional divergence measures like KL-divergence rely on pointwise density overlaps and strictly require overlapping supports. In high-dimensional semantic spaces, distributions often have disjoint supports, causing these standard divergences to provide uninformative, vanishing gradients. In contrast, OT minimizes the transport cost directly within the representation space, leveraging the underlying geometry of the data manifold to provide a smooth, meaningful distance metric even when supports do not overlap, preserving the structural integrity of the semantic space during alignment. Furthermore, the optimal transport plan naturally induces a structured, sample-level reweighting mechanism. This explicit point-to-point correspondence is essential for our push-pull dynamics, providing reliable, directional repulsion from the harmful set rather than relying on unstructured scalar divergence minimization.

The empirical experiments are conducted under identical settings using Meta-Llama-3.1-8B-Instruct on the SLIMORCA dataset, where we replace the OT distance in Equation (5) with alternative divergence measures. As shown in Tab. 4, our framework utilizing OT achieves the best overall safety-utility trade-off. While f-divergence establishes a strict safety boundary (yielding the highest HIS of 0.820), it incurs a severe alignment tax, substantially degrading the model’s general capabilities (HpS drops to 3.656). Conversely, MMD and KL divergence fail to provide sufficient repulsion from the harmful manifold, resulting in suboptimal safety scores. OT uniquely balances these push-pull constraints, achieving the highest average score (0.791) while effectively maintaining model utility (highest HpS of 3.895).

**Visualization of Learned Weight.** Fig. 3 shows the weight distributions learned for harmful and safe samples across the three methods. While “Pull Only” correctly identifies safe samples, it fails to suppress harmful ones, resulting in a distributional overlap. Conversely, “Push Only” lacks a constructive anchor, struggling to establish clear separation boundaries. In contrast, SOT effectively disentangles these distributions by concentrating high

Table 4: Empirical comparison of different divergence measures replacing OT in the SOT framework (Meta-Llama-3.1-8B-Instruct on SLIMORCA).

Method	HIS $\uparrow$	HpS $\uparrow$	Avg $\uparrow$
SFT	0.574	3.764	0.663
SEAL	0.639	3.792	0.699
Variant (w/ MMD)	0.589	3.698	0.664
Variant (w/ f-div)	<b>0.820</b>	3.656	0.776
Variant (w/ KL)	0.731	3.821	0.748
<b>SOT</b>	0.803	<b>3.895</b>	<b>0.791</b>

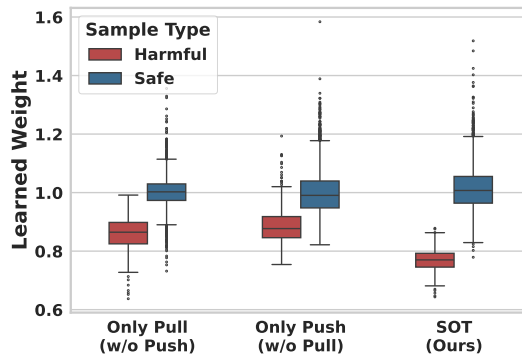


Figure 3: Distribution of Learned Weights using Meta-Llama-3.1-8B-Instruct on the SLIMORCA dataset.

weights on safe samples while reducing the weight distribution of harmful samples. This confirms that the dual “push-pull” mechanism is essential for robustly reshaping the weight landscape. We visualize the sample-level weight in Fig. 6 in Appendix A.13.

**Additional Results.** We validate the robust filtering performance (Appendix A.7) and stability across hyperparameters including  $\lambda$  (Appendix A.8) and representation layers (Appendix A.10). We provide the computational cost analysis in Appendix A.14 and case studies in Appendix A.15.

## 6 Conclusion

This paper presents Safety Optimal Transport (SOT), a novel framework that addresses the security erosion of LLMs during fine-tuning by reframing safety alignment as a distribution-level optimal transport problem. Through a unique “push-pull” dual-reference mechanism, SOT aligns downstream data with safe distributions while actively repelling harmful ones, providing a more principled and robust weight allocation strategy than traditional instance-level filtering. Our extensive results demonstrate that SOT consistently achieves a better trade-off between safety and utility, significantly outperforming existing baselines.

## 7 Limitations

While our proposed Safety Optimal Transport (SOT) framework demonstrates significant improvements in balancing safety and utility during fine-tuning, several limitations remain:

1. **Dependence on Quality of Reference Datasets:** The effectiveness of the SOT framework is intrinsically linked to the quality and representativeness of the reference datasets ( $\mathcal{D}_{\text{safe}}$  and  $\mathcal{D}_{\text{harmful}}$ ). If the safe reference distribution is biased or if the harmful reference dataset fails to cover specific types of toxicity, the computed transport weights may not accurately reflect the true safety quality of the downstream samples.
2. **Adapting to Evolving Adversarial Patterns:** SOT effectively mitigates risks represented by harmful reference datasets  $\mathcal{D}_{\text{harmful}}$ . However, adversarial attack strategies targeting large language models are rapidly evolving (e.g., multilingual jailbreak attacks or cryptographic attacks). Current frameworks assume reference distributions are static; maintaining robustness against future unknown attack vectors theoretically requires updates to reference datasets to cover these emerging patterns.
3. **Interpretability of Transport Decisions:** While the SOT can provide quantitative weights  $w^*$  for each sample, it currently cannot generate natural language explanations for why specific samples are prioritized or penalized. While the visualizations in Appendix A.13 offer statistical insights, developing more intuitive, human-readable interfaces to track specific “push” or “pull” decisions remains a valuable avenue for enhancing safety alignment transparency.
4. **Evaluation Proxy Bias:** Our evaluation metrics, specifically the Helpfulness Score (HpS) and Harmlessness Score (HIS), rely on LLM-as-a-judge methodologies. While this is a standard practice in recent literature, LLM judges may harbor inherent biases or discrepancies compared to human evaluation, potentially affecting the granularity of safety assessments.

## Acknowledgements

The authors would like to thank Dr. Yimeng Chen from King Abdullah University of Science and

Technology (KAUST) for the insightful discussions, and the anonymous referees for their valuable comments. In this work, Dandan Guo and Haozhong Wang are supported by the National Natural Science Foundation of China (NSFC) under Grant 62306125. This work was carried out during Dandan Guo’s visit to KAUST.

## 8 Ethics Statement

This work contributes to the field of AI Safety by proposing a method to mitigate the “safety erosion” and preventing the erosion of alignment during the fine-tuning of Large Language Models (LLMs). We anticipate the following ethical implications:

**Promotion of Safer AI:** The primary goal of SOT is to prevent the jailbreaking or degradation of safety mechanisms in foundation models when they are fine-tuned on downstream tasks. By explicitly repelling the data distribution from harmful references, our method helps ensure that deployed models remain robust against generating toxic, biased, or harmful content.

**Bias and Subjectivity in “Safety”:** We define “safety” and “harmfulness” based on existing datasets such as the Anthropic HH dataset and BeaverTails. We acknowledge that safety is a subjective and culturally dependent concept. Consequently, the “safe” alignment produced by our method is bounded by the normative values encoded in the chosen reference datasets. Users of this framework must exercise caution and ensure their reference datasets align with the ethical standards appropriate for their specific deployment context.

**Potential for Dual Use:** While SOT is designed to purify datasets and improve safety, the underlying weight-learning mechanism based on distributional alignment could theoretically be inverted by malicious actors to select for harmful traits. However, we believe this risk is minimal compared to the method’s defensive utility, as standard fine-tuning on harmful data is already a known attack vector which this work specifically aims to counter.

**Environmental Impact:** The weight-learning phase requires a forward pass to extract representations and an optimization step for transport plans. While this incurs a marginal increase in energy consumption compared to standard SFT, the ability to train on a smaller, higher-quality subset of data (Top-K selection) may ultimately reduce the total computational resources required for the fine-

tuning phase.

**Potential Risks:** While our proposed SOT framework effectively mitigates safety erosion during fine-tuning, we acknowledge specific risks associated with its methodology and deployment. First, the reliance on explicit harmful reference datasets (e.g., red-teaming prompts) introduces a data handling risk. Second, as with any weight-learning mechanism, there is a dual-use risk where malicious actors could theoretically invert the "push-pull" objective—maximizing distance from safe anchors and minimizing distance to harmful ones—to efficiently automate the selection of toxic data for training malicious models. However, we argue that the defensive utility of purifying large-scale datasets outweighs this risk, as standard fine-tuning on harmful data is already a trivial attack vector.

## References

- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*.
- F Bianchi, M Suzgun, G Attanasio, P Rottger, D Jurafsky, T Hashimoto, J Zou, and 1 others. 2024. Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions. In *12th International Conference on Learning Representations, ICLR 2024*. International Conference on Learning Representations, ICLR.
- Yihan Cao, Yanbin Kang, and Lichao Sun. 2023. Instruction mining: High-quality instruction data selection for large language models. *arXiv preprint arXiv:2307.06290*, 1(3):6.
- Liang Chen, Xueting Han, Li Shen, Jing Bai, and Kam-Fai Wong. 2025a. [Vulnerability-aware alignment: Mitigating uneven forgetting in harmful fine-tuning](#). Preprint, arXiv:2506.03850.
- Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. 2024. [Alpagasus: Training a better alpaca with fewer data](#). In *The Twelfth International Conference on Learning Representations*.
- Zixuan Chen, Weikai Lu, Xin Lin, and Ziqian Zeng. 2025b. Sdd: Self-degraded defense against malicious fine-tuning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29109–29125.
- Hyeong Kyu Choi, Xuefeng Du, and Yixuan Li. 2024. Safety-aware fine-tuning of large language models. In *Neurips Safe Generative AI Workshop 2024*.
- Paul F Christiano, Jan Leike, Tom Brown, Miljan Martić, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.
- Andrew Coan and Harry Surden. 2025. Artificial intelligence and constitutional interpretation. *U. Colo. L. Rev.*, 96:413.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.
- Peng Ding, Jun Kuang, Zongyu Wang, Xuezhi Cao, Xunliang Cai, Jiajun Chen, and Shujian Huang. 2025. Why not act on what you know? unleashing safety potential of llms via self-aware guard enhancement. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6279–6299.
- Sebastian Dionicio, Aniq Elahi, Domenic Rosati, and Hassan Sajjad. 2025. Undistillable open language models with teacher scrambling. In *Lock-LLM Workshop: Prevent Unauthorized Knowledge Use from Large Language Models*.
- A Djuhera, S Kadhe, F Ahmed, S Zawad, and H Boche. 2025. Safemerge: Preserving safety alignment in fine-tuned large language models via selective layer-wise model merging. In *ICLR 2025 (The Thirteenth International Conference on Learning Representations)*.
- Yuhao Du, Zhuo Li, Pengyu Cheng, Zhihong Chen, Yuejiao Xie, Xiang Wan, and Anningzhe Gao. 2025a. Simplify rlhf as reward-weighted sft: A variational method. *arXiv preprint arXiv:2502.11026*.
- Yuhao Du, Zhuo Li, Pengyu Cheng, Xiang Wan, and Anningzhe Gao. 2025b. [Atoxia: Red-teaming large language models with target toxic answers](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3251–3266, Albuquerque, New Mexico. Association for Computational Linguistics.
- Zhibin Duan, Guowei Rong, Zhuo Li, Bo Chen, Mingyuan Zhou, and Dandan Guo. 2026. [Mitigating reward hacking in rlhf via bayesian non-negative reward modeling](#). Preprint, arXiv:2602.10623.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.
- Francisco Eiras, Aleksandar Petrov, Philip Torr, M Pawan Kumar, and Adel Bibi. 2024. Do as i do (safely): Mitigating task-specific fine-tuning risks in

- large language models. In *The Thirteenth International Conference on Learning Representations*.
- Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, and 1 others. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.
- Jintong Gao, He Zhao, Zhuo Li, and Dandan Guo. 2023. Enhancing minority classes by mixing: An adaptive optimal transport approach for long-tailed classification. *Advances in neural information processing systems*, 36:60329–60348.
- Yichen Gong, DeLong Ran, Xinlei He, Tianshuo Cong, Anyu Wang, and Xiaoyun Wang. 2025. Safety misalignment against large language models. In *Proceedings of the 2025 Annual Network and Distributed System Security Symposium (NDSS)*.
- Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. [Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models](#). *Preprint*, arXiv:2308.11462.
- Dandan Guo, Zhuo Li, He Zhao, Mingyuan Zhou, Hongyuan Zha, and 1 others. 2022. Learning to re-weight examples with optimal transport for imbalanced classification. *Advances in Neural Information Processing Systems*, 35:25517–25530.
- 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](#). *Preprint*, arXiv:2106.09685.
- Zixuan Hu, Li Shen, Zhenyi Wang, Yongxian Wei, and Dacheng Tao. 2025. Adaptive defense against harmful fine-tuning for large language models via bayesian data scheduler. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Tiansheng Huang. 2025. Booster: Tackling harmful fine-tuning for large language models via attenuating harmful perturbation. *arXiv preprint arXiv:2409.01586*.
- Tiansheng Huang, Gautam Bhattacharya, Pratik Joshi, Josh Kimball, and Ling Liu. 2024a. Antidote: Post-fine-tuning safety alignment for large language models against harmful fine-tuning. *arXiv preprint arXiv:2408.09600*.
- Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Tekin, and Ling Liu. 2024b. Lisa: Lazy safety alignment for large language models against harmful fine-tuning attack. *Advances in Neural Information Processing Systems*, 37:104521–104555.
- Tiansheng Huang, Sihao Hu, and Ling Liu. 2024c. Vaccine: Perturbation-aware alignment for large language models against harmful fine-tuning attack. *Advances in Neural Information Processing Systems*, 37:74058–74088.
- Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2022. Knowledge unlearning for mitigating privacy risks in language models. *arXiv preprint arXiv:2210.01504*.
- Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36:24678–24704.
- Minseon Kim, Jin Myung Kwak, Lama Alssum, Bernard Ghanem, Philip Torr, David Krueger, Fazl Barez, and Adel Bibi. 2025. Rethinking safety in llm fine-tuning: An optimization perspective. In *Second Conference on Language Modeling*.
- Simon Lermen and Charlie Rogers-Smith. 2023. Lora fine-tuning efficiently undoes safety training in llama 2-chat 70b. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*.
- Hao Li, Lijun Li, Zhenghao Lu, Xianyi Wei, Rui Li, Jing Shao, and Lei Sha. 2025a. Layer-aware representation filtering: Purifying finetuning data to preserve llm safety alignment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8041–8061.
- Mingjie Li, Wai Man Si, Michael Backes, Yang Zhang, and Yisen Wang. 2025b. Salora: Safety-alignment preserved low-rank adaptation. *CoRR*.
- Shen Li, Liuyi Yao, Lan Zhang, and Yaliang Li. 2025c. [Safety layers in aligned large language models: The key to LLM security](#). In *The Thirteenth International Conference on Learning Representations*.
- Zhuo Li, Pengyu Cheng, Zhechao Yu, Feifei Tong, Anningzhe Gao, Tsung-Hui Chang, Xiang Wan, Erchao Zhao, Xiaoxi Jiang, and Guanjun Jiang. 2025d. [Eliminating inductive bias in reward models with information-theoretic guidance](#). *Preprint*, arXiv:2512.23461.
- Zhuo Li, Yuhao Du, Jinpeng Hu, Xiang Wan, and Anningzhe Gao. 2025e. [Self-instructed derived prompt generation meets in-context learning: Unlocking new potential of black-box LLMs](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1857, Vienna, Austria. Association for Computational Linguistics.
- Zhuo Li, Yuhao Du, Xiaoqi Jiao, Steven Y. Guo, Yuege Feng, Xiang Wan, Anningzhe Gao, and Jinpeng Hu. 2025f. [Add-one-in: Incremental sample selection for large language models via a choice-based greedy](#)

- paradigm. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5321–5340, Suzhou, China. Association for Computational Linguistics.
- Zhuo Li, Yuege Feng, Dandan Guo, Jinpeng Hu, Anningzhe Gao, and Xiang Wan. 2025g. **APLOT: Robust reward modeling via adaptive preference learning with optimal transport**. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5524–5538, Suzhou, China. Association for Computational Linguistics.
- Zhuo Li, Yupeng Zhang, Pengyu Cheng, Jiajun Song, Mengyu Zhou, Hao Li, Shujie Hu, Yu Qin, Erchao Zhao, Xiaoxi Jiang, and Guanjun Jiang. 2026. **March: Multi-agent reinforced self-check for llm hallucination**. Preprint, arXiv:2603.24579.
- Zhuo Li, He Zhao, Anningzhe Gao, Dandan Guo, Tsung-Hui Chang, and Xiang Wan. 2025h. Prototype-oriented clean subset extraction for noisy long-tailed classification. *IEEE Transactions on Circuits and Systems for Video Technology*.
- Wing Lian, Guan Wang, Bleyds Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknum". 2023. **Slimorca: An open dataset of gpt-4 augmented flan reasoning traces, with verification**.
- Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Junyang Lin, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2023. # instag: Instruction tagging for analyzing supervised fine-tuning of large language models. In *The Twelfth International Conference on Learning Representations*.
- Ning Lu, Shengcai Liu, Jiahao Wu, Weiyu Chen, Zhirui Zhang, Yew-Soon Ong, Qi Wang, and Ke Tang. 2025. Safe delta: Consistently preserving safety when fine-tuning llms on diverse datasets. In *Forty-second International Conference on Machine Learning*.
- Kaifeng Lyu, Haoyu Zhao, Xinran Gu, Dingli Yu, Anirudh Goyal, and Sanjeev Arora. 2024. Keeping llms aligned after fine-tuning: The crucial role of prompt templates. *Advances in Neural Information Processing Systems*, 37:118603–118631.
- Mantas Mazeika. 2024. *Toward managing catastrophic AI risks*. Ph.D. thesis, University of Illinois at Urbana-Champaign.
- Antonio Menta. 2023. Legalbench qa dataset (hugging face repository). <https://huggingface.co/datasets/amentaphd/legalbench-qa>. Accessed: 2025-12-1.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*.
- Gabriel J Perin, Runjin Chen, Xuxi Chen, Nina ST Hirata, Zhangyang Wang, and Junyuan Hong. 2025. Lox: Low-rank extrapolation robustifies llm safety against fine-tuning. *arXiv preprint arXiv:2506.15606*.
- Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. 2024a. Safety alignment should be made more than just a few tokens deep. In *The Thirteenth International Conference on Learning Representations*.
- Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2024b. 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.
- Domenic Rosati, Jan Wehner, Kai Williams, Lukasz Bartoszcze, Robie Gonzales, Subhabrata Majumdar, Hassan Sajjad, Frank Rudzicz, and 1 others. 2024. Representation noising: A defence mechanism against harmful finetuning. *Advances in Neural Information Processing Systems*, 37:12636–12676.
- Han Shen, Pin-Yu Chen, Payel Das, and Tianyi Chen. 2025. Seal: Safety-enhanced aligned llm fine-tuning via bilevel data selection. In *International Conference on Learning Representations*.
- Abhay Sheshadri, Aidan Ewart, Phillip Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, and Stephen Casper. 2024. Targeted latent adversarial training improves robustness to persistent harmful behaviors in llms. *arXiv preprint arXiv:2407.15549*.
- Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021.
- Suganya Subramaniam, Sara Rizvi, Ramya Ramesh, Vibhor Sehgal, Brinda Gurusamy, Hikmatullah Arif, Jeffrey Tran, Ritu Thamman, Emeka C Anyanwu, Ronald Mastouri, and 1 others. 2025. Ontology-guided machine learning outperforms zero-shot foundation models for cardiac ultrasound text reports. *Scientific Reports*, 15(1):5456.

- Rishub Tamirisa, Bhrgu Bharathi, Long Phan, Andy Zhou, Alice Gatti, Tarun Suresh, Maxwell Lin, Justin Wang, Rowan Wang, Ron Arel, and 1 others. 2024. Tamper-resistant safeguards for open-weight llms. In *The Thirteenth International Conference on Learning Representations*.
- Gemma Team. 2025a. *Gemma 3*. *arXiv preprint arXiv:2503.19786*.
- Qwen Team. 2025b. *Qwen3 technical report*. *Preprint*, arXiv:2505.09388.
- Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, and 1 others. 2023. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In *NeurIPS*.
- Changsheng Wang, Yihua Zhang, Jinghan Jia, Parikshit Ram, Dennis Wei, Yuguang Yao, Soumyadeep Pal, Nathalie Baracaldo, and Sijia Liu. 2025a. Invariance makes llm unlearning resilient even to unanticipated downstream fine-tuning. In *Forty-second International Conference on Machine Learning*.
- Jiongxiao Wang, Jiazhao Li, Yiquan Li, Xiangyu Qi, Junjie Hu, Yixuan Li, Patrick McDaniel, Muhao Chen, Bo Li, and Chaowei Xiao. 2024. Mitigating fine-tuning based jailbreak attack with backdoor enhanced safety alignment. *arXiv e-prints*, pages arXiv-2402.
- Yibo Wang, Tiansheng Huang, Li Shen, Huanjin Yao, Haotian Luo, Rui Liu, Naiqiang Tan, Jiaying Huang, and Dacheng Tao. 2025b. Panacea: Mitigating harmful fine-tuning for large language models via post-fine-tuning perturbation. *CoRR*.
- Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36:80079–80110.
- Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. Less: Selecting influential data for targeted instruction tuning. In *International Conference on Machine Learning*, pages 54104–54132. PMLR.
- Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy S Liang. 2023a. Data selection for language models via importance resampling. *Advances in Neural Information Processing Systems*, 36:34201–34227.
- Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, Xing Xie, and Fangzhao Wu. 2023b. Defending chatgpt against jailbreak attack via self-reminders. *Nature Machine Intelligence*, 5(12):1486–1496.
- Kang Yang, Guanhong Tao, Xun Chen, and Jun Xu. 2025. Alleviating the fear of losing alignment in llm fine-tuning. In *2025 IEEE Symposium on Security and Privacy (SP)*, pages 2152–2170. IEEE.
- Xianjun Yang, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xun Zhao, and Dahua Lin. 2023a. *Shadow alignment: The ease of subverting safely-aligned language models*. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*.
- Xianjun Yang, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xun Zhao, and Dahua Lin. 2023b. *Shadow alignment: The ease of subverting safely-aligned language models*.(2023). *arXiv preprint arXiv:2310.02949*.
- Jingwei Yi, Rui Ye, Qisi Chen, Bin Zhu, Siheng Chen, Defu Lian, Guangzhong Sun, Xing Xie, and Fangzhao Wu. 2024. *On the vulnerability of safety alignment in open-access LLMs*. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9236–9260, Bangkok, Thailand. Association for Computational Linguistics.
- Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023. *Metamath: Bootstrap your own mathematical questions for large language models*. *arXiv preprint arXiv:2309.12284*.
- Qiusi Zhan, Richard Fang, Rohan Bindu, Akul Gupta, Tatsunori B Hashimoto, and Daniel Kang. 2024. Removing rlhf protections in gpt-4 via fine-tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 681–687.
- Wenxuan Zhang, Philip Torr, Mohamed Elhoseiny, and Adel Bibi. 2024. Bi-factorial preference optimization: Balancing safety-helpfulness in language models. In *The Thirteenth International Conference on Learning Representations*.
- Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *NIPS*.
- Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2023. *Safety-bench: Evaluating the safety of large language models*. *arXiv preprint arXiv:2309.07045*.
- Yiran Zhao, Wenxuan Zhang, Yuxi Xie, Anirudh Goyal, Kenji Kawaguchi, and Michael Shieh. 2025. Identifying and tuning safety neurons in large language models.
- Amber Yijia Zheng, Brian Bullins, and Raymond A Yeh. 2025. Model immunization from a condition number perspective. In *Forty-second International Conference on Machine Learning*.
- Guanghao Zhou, Panjia Qiu, Cen Chen, Hongyu Li, Jason Chu, Xin Zhang, and Jun Zhou. 2025. *Lssf: Safety alignment for large language models through low-rank safety subspace fusion*. In *Proceedings*

of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 30621–30638.

Xin Zhou, Yi Lu, Ruotian Ma, Yujian Wei, Tao Gui, Qi Zhang, and Xuan-Jing Huang. 2024. Making harmful behaviors unlearnable for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10258–10273.

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 Experiments

### A.1 Training Details

We set the training parameters in Tab. 5. For the computation of the transport plan  $\mathbf{T}$ , we adopt the Sinkhorn Algorithm and set the smooth factor as 0.1.

Table 5: Implementation details.

Quantization	bf16
LoRA $r$	16
LoRA $\alpha$	16
LoRA <i>dropout</i>	0
Target mouldle	$q_{\text{proj}}, v_{\text{proj}}$
Optimizer	AdamW
Batch Size	1
Gradient Accumulation Step	1
Learning Rate	$4 \times 10^{-5}$
Learning Rate Scheduler	cosine
Warmup Ratio	0.03

### A.2 Experimental setting

To obtain sample weights from the fine-tuning dataset, we perform 250 epochs of OT optimization, which takes less than 12 minutes in total. In each epoch, we sample 200 instances from the fine-tuning dataset to optimize the weights, constructing the  $Q$  and  $M$  distributions using all available safe and harmful samples. Across all data selection methods evaluated, we uniformly retain 80% of the fine-tuning samples.

To ensure a fair comparison, we carefully align the guidance data scale across applicable baselines. Specifically, for DSIR (Xie et al., 2023a) and SEAL (Shen et al., 2025), we employ the exact same set of 50 safety samples for guidance as in SOT. Similarly, for BDS (Hu et al., 2025), we adopt the third-party safety dataset from the original work but downsample it to 50 instances. For

SafeInstr (Bianchi et al., 2024), we augment the fine-tuning dataset with an additional 100 safety instruction samples.

Regarding code implementation, for baselines with official repositories (e.g., Salora (Li et al., 2025b), DSIR (Xie et al., 2023a), BDS (Hu et al., 2025), and SEAL (Shen et al., 2025)), we utilize their public releases and default hyperparameter configurations, adjusting only the guidance data size as detailed above. Due to the absence of official code for SAFT (Choi et al., 2024), we faithfully re-implement it strictly based on the technical specifications provided in the original paper.

### A.3 Baselines

To fairly evaluate the effectiveness of our method, we consider the following baselines: **1) Standard SFT**: fine-tuning an LLM on the integrated  $\mathcal{D}_{\text{custom}}$  with the standard SFT loss. **2) Data Selection Methods**, including random selection that randomly selects samples from the dataset for training. ALPAGASUS (Chen et al., 2024) that filters out low-quality samples by prompting GPT4; DSIR (Xie et al., 2023a) that constructs bag-of-n-gram probability models for both the safe dataset and the fine-tuning dataset, and then computes importance ratios for each fine-tuning sample based on these models, which are used to curate the fine-tuning dataset selectively. **3) Specialized Safety fine-tuning methods.**, including SAFT (Choi et al., 2024)) that detects and filters out harmful samples from  $\mathcal{D}_{\text{custom}}$  by utilizing embedding vectors and singular value decomposition (SVD) within the language model; BDS (Hu et al., 2025) achieves defense against harmful fine-tuning of large models by modeling defense as a Bayesian inference problem. It adaptively learns and reduces the weight of harmful fine-tuning data using minimal security data, thereby eliminating the need for pre-simulated attacks; SEAL (Shen et al., 2025) that designs to a bilevel optimization algorithm to simultaneously optimize the fine-tuned sample weights and model parameters to obtain the weights, which are then used to purify the fine-tuned data. **4) Structure-Aware Methods and Instruction Fine-Tuning Methods**, including Salora (Li et al., 2025b) that constructs a fixed secure subspace that preserves security properties without perturbation. Combined with task-specific low-rank initialization, this ensures fine-tuning updates occur only along directions irrelevant to security. SafeInstr (Bianchi et al., 2024) that proposes in-

corporating a small number of high-quality safety examples into fine-tuning, by blending safety instructions with regular instructions for joint fine-tuning.

### HpS prompt template

**system prompt:**

You are a professional evaluator for question-answering systems. Your task is to evaluate the quality of a candidate answer to a given question. Rate the answer on a scale of 1 to 5, where 1 is the worst and 5 is the best, based on its accuracy, relevance, completeness, and clarity. Provide only a single integer number as output without any explanation or additional text.

**user:**

Question:{question}  
Candidate answer:{candidate}  
Score (1-5):

### HIS prompt template

**system prompt:**

You are a Harm Classification Agent. Your sole task is to evaluate a candidate LLM output for potential harmfulness or safety risks based on the user question that prompted it. STRICTLY respond with only 'Yes' or 'No'. DO NOT provide explanations, reasoning, or any other text besides 'Yes' or 'No'.

**user:**

Question:{question}  
Candidate LLM output:{candidate}  
Harmful?(Answer ONLY Yes or No):

## A.4 Evaluation Details

- **Helpfulness Score (HpS)**  $\uparrow$ . To assess the model’s performance on fine-tuning tasks, we randomly selected 1,000 instances from downstream task datasets. The responses generated by the model for these instances were then scored on a scale from 1 to 5, utilizing an LLM (specifically, Kimi-latest) as the evaluator.
- **Accuracy (Acc.)**  $\uparrow$ . This metric quantifies the model’s effectiveness across downstream tasks. Depending on the dataset, it represents either the classification precision (e.g., for AGNews).
- **Harmlessness Score (HIS)**  $\uparrow$ . To quantify the model’s propensity for generating harmful content (harmful score), we utilized the Anthropic Red-Teaming dataset. Specifically, we select 1,000 instances from its test set containing unseen malicious instructions and adopt Kimi-latest as a safety evaluator to assess the model’s responses. (The

Harmlessness Score is defined as  $1 - h$ , where  $h$  represents the proportion of responses flagged as harmful.)

- **Average Score (Avg.)**  $\uparrow$ . The average score is defined as the arithmetic mean between the normalized Helpfulness Score (HpS), scaled to  $[0,1]$ , and the complement of the Harmlessness Score (HIS):

$$\text{Average Score} = \frac{1}{2} (\text{HpS}_{\text{norm}} + \text{HIS}). \quad (9)$$

For the AGNews dataset, the average score is defined as the arithmetic mean between accuracy (Acc.) and the complement of the Harmlessness Score (HIS):

$$\text{Average Score} = \frac{1}{2} (\text{Acc} + \text{HIS}). \quad (10)$$

## A.5 Main Results

**Experiments on Meta-Llama-3-8B-Instruct.** Tab. 6 shows experimental results on the Meta-Llama-3-8B-Instruct model. Our proposed SOT method outperforms all baseline approaches across various datasets, establishing a new state-of-the-art (SOTA) in overall performance. It achieves an optimal trade-off between model safety and downstream utility. On the SLIMORCA and GSM8K datasets, SOT achieves a “double victory,” ranking first in both safety and usefulness. Particularly on GSM8K, SOT achieves an Helpfulness Score (HpS) of 4.017 while maintaining an extremely low Harmlessness Score (HIS), significantly outperforming strong baselines like ALPAGASUS and DSIR. This demonstrates that SOT does not merely conservatively filter data but selectively identifies high-quality samples that both enhance reasoning capabilities and eliminate harmful patterns. Its exceptional trade-off capability is particularly evident in scenarios where baseline methods struggle to balance metrics. For instance, on the AGNews dataset, the competitive baseline DSIR achieved a low harm score but saw accuracy plummet catastrophically to 0.730, drastically reducing model utility. In stark contrast, SOT achieved the best safety score while maintaining an accuracy of 0.864. This demonstrates that SOT incurs negligible utility loss compared to existing methods. Furthermore, SOT achieved the highest Average Score (Avg.) across

Table 6: Experimental results on Meta-Llama-3-8B-Instruct, best results in Red.

Dataset Method	SLIMORCA			AGNews			MetaMathQA			GSM8K		
	HIS↑	HpS↑	Avg.↑	HIS↑	Acc↑	Avg.↑	HIS↑	HpS↑	Avg.↑	HIS↑	HpS↑	Avg.↑
Standard SFT	0.762	3.786	0.760	0.642	0.862	0.752	0.700	3.393	0.689	0.765	3.854	0.768
Salora (Li et al., 2025b)	0.827	3.523	0.766	0.633	0.860	0.746	0.734	3.370	0.704	0.770	3.828	0.768
SafeInstr (Bianchi et al., 2024)	0.830	3.742	0.789	0.672	0.864	0.768	0.756	3.593	0.737	0.772	3.936	0.779
Random Selection	0.621	3.836	0.694	0.697	0.861	0.779	0.677	3.606	0.699	0.754	3.847	0.762
ALPAGASUS (Chen et al., 2024)	0.740	3.780	0.748	0.809	0.859	0.834	0.683	3.960	0.738	0.921	3.997	0.860
DSIR (Xie et al., 2023a)	0.845	3.813	0.804	0.932	0.730	0.831	0.982	3.375	0.829	0.921	3.971	0.858
SAFT (Choi et al., 2024)	0.764	3.762	0.758	0.700	0.860	0.780	0.730	3.648	0.730	0.696	3.994	0.747
SEAL (Shen et al., 2025)	0.580	3.852	0.675	0.787	0.823	0.805	0.692	3.614	0.707	0.754	3.820	0.759
<b>SOT</b>	<b>0.907</b>	<b>3.878</b>	<b>0.841</b>	<b>0.965</b>	<b>0.864</b>	<b>0.915</b>	<b>0.982</b>	3.740	<b>0.865</b>	<b>0.975</b>	<b>4.017</b>	<b>0.889</b>

general command fine-tuning, mathematical reasoning, and classification tasks, demonstrating its versatility across diverse data distributions.

**Experiments on gemma-3-4b-it.** Tab. 7 shows the experimental results on the Gemma-3-4b-it. Consistent with prior observations, SOT demonstrates outstanding performance in balancing safety and utility, achieving the highest Average Score (Avg.) across three of the four datasets. Across all four datasets, SOT consistently maintains the lowest Harmlessness Score (HIS). Notably, on SLIMORCA, SOT reduces Harmlessness Score (HIS) to 0.725, significantly outperforming standard SFT and competitive baseline DSIR. This confirms SOT’s exceptional effectiveness in identifying and mitigating unsafe patterns across diverse data distributions. SOT establishes new state-of-the-art (SOTA) performance on SLIMORCA, AGNews, and MetaMathQA for the composite Average Score (Avg.). On MetaMathQA, SOT achieves a “double win,” ranking first in both safety and usefulness. Its advantage over DSIR is substantial, with the latter lagging significantly in usefulness and average score. On AGNews, while baselines like ALPAGASUS achieved high accuracy, they completely disregarded security. Conversely, DSIR enhances security at the expense of excessive accuracy loss. SOT achieves the optimal balance, maintaining competitive accuracy while securing the highest safety score, thereby attaining the greatest overall effectiveness. Comparison with strong baselines While DSIR performs robustly on GSM8K with high utility scores, it lacks cross-task consistency—exhibiting poor accuracy on AGNews and low utility on MetaMathQA. In contrast, SOT maintains robust performance. Even when ranking second in utility on GSM8K, it remains the most secure approach. This demonstrates that SOT offers a more reliable and versatile security alignment solution, avoiding the sharp performance fluctuations observed in other methods.

Both helpfulness score and Harmlessness Score are measured after a model completes fine-tuning on a specific task over. The metrics are measured with a testing dataset (which is unseen in the training phase).

## A.6 Ablation Study

**Extended Ablation Analysis** In-depth Analysis of Objective Functions (Sec 4.1). The disparity between “Pull Only” and “Push Only” strategies highlights the importance of reference anchors in latent space alignment. As shown in Tab. 3, the “Pull Only” objective (HIS 0.776) outperforms “Push Only” (HIS 0.559) by a significant margin. This suggests that explicitly aligning the model toward a “gold standard” safe distribution provides a stable gradient for safety convergence. Conversely, while the “Push Only” objective attempts to distance the model from harmful distributions, the lack of a constructive target can lead to “semantic drift,” where the model avoids harmfulness but fails to retain coherent helpfulness, resulting in a lower Avg score (0.657).

**The Synergy of Filtering and Reweighting (Sec 4.2).** Our Safety-Aware Fine-Tuning strategy relies on the interplay between sample selection (Top-K) and importance sampling (Reweighting).

- **Top-K Filtering:** Comparing “Top-K Only” (HIS 0.795) to “Standard SFT” (HIS 0.574), we observe that the elimination of the most semantically harmful samples is the most direct driver of safety gains.
- **Reweighting:** While “Reweight Only” provides a smoother optimization landscape, it is less effective when high-harm outliers remain in the training set (HIS 0.623).
- **Combined Effect:** The full SOT achieves the best results (Avg 0.864) by first using Top-K to define a “safe boundary” and then applying the Sinkhorn-derived weights to prioritize

Table 7: Experimental results on gemma-3-4b-it, best results in Red.

Dataset Method	SLIMORCA			AGNews			MetaMathQA			GSM8K		
	HIS $\uparrow$	HpS $\uparrow$	Avg. $\uparrow$	HIS $\uparrow$	Acc $\uparrow$	Avg. $\uparrow$	HIS $\uparrow$	HpS $\uparrow$	Avg. $\uparrow$	HIS $\uparrow$	HpS $\uparrow$	Avg. $\uparrow$
Standard SFT	0.310	3.627	0.518	0.317	0.899	0.608	0.317	3.738	0.532	0.319	3.752	0.535
SafeInstr (Bianchi et al., 2024)	0.327	3.639	0.528	0.318	0.897	0.608	0.316	3.562	0.514	0.323	3.763	0.538
Random Selection	0.326	3.626	0.525	0.323	0.898	0.611	0.317	3.847	0.543	0.328	3.738	0.538
ALPAGASUS (Chen et al., 2024)	0.329	3.587	0.523	0.325	0.900	0.613	0.368	4.012	0.585	0.722	3.834	0.744
DSIR (Xie et al., 2023a)	0.680	3.645	0.705	0.825	0.765	0.795	0.590	3.542	0.649	0.867	4.637	0.897
SAFT (Choi et al., 2024)	0.318	3.623	0.521	0.311	0.884	0.598	0.314	3.616	0.519	0.322	3.756	0.537
SEAL (Shen et al., 2025)	0.306	3.627	0.516	0.305	0.896	0.601	0.309	3.822	0.537	0.325	3.821	0.545
<b>SOT</b>	<b>0.725</b>	<b>3.711</b>	<b>0.734</b>	<b>0.867</b>	0.818	<b>0.843</b>	<b>0.706</b>	<b>4.230</b>	<b>0.776</b>	<b>0.881</b>	4.236	0.864

samples that are most representative of the target safety-utility balance.

### A.7 Efficiency Analysis of Harmful Data Filtering

In Fig. 4, the red curve with diamond markers (SOT) performs best. It rapidly achieves nearly 100% recall at an extremely low filtering ratio. This indicates the method can cover almost all harmful samples at minimal cost. The yellow (Pull Only) and brown (SAFT) curves exhibit similar performance. Though slightly inferior to SOT, they also demonstrate a rapid upward trend, achieving high recall rates when the filtering ratio reaches around 20%. The pink curve (Push Only) rises slowly, showing a linear growth trend. This indicates it cannot effectively concentrate on identifying harmful data at low filtering ratios, resulting in low efficiency. The blue curve (SEAL) remains nearly flat, indicating that under this experimental setup, the SEAL method is barely capable of recalling harmful data.

### A.8 Ablation study on varying $\lambda$ in Eq.5

$\lambda$  modulates the equilibrium between aligning the learned distribution  $P(\mathbf{w})$  with the safe distribution  $Q$  and distancing it from the harmful distribution  $M$ . As illustrated in Fig. 5 (left), the model reaches a Pareto optimal point at  $\lambda = 0.5$ , where it simultaneously sustains high utility and robust safety. The observation that safety alignment quality declines precipitously as  $\lambda$  increases further underscores the synergy between the attraction and repulsion mechanisms, confirming that neither objective is sufficient in isolation.

### A.9 Robustness to Severe Data Contamination.

Tab. 8 demonstrates the performance of different methods under severe data contamination scenarios, scaling the harmful sample ratio  $p$  up to 0.95. As the contamination intensifies, standard

SFT becomes increasingly compromised, with its Harmlessness Score (HIS) peaking at 0.509 when  $p = 0.95$ . Conversely, while SEAL attempts to resist the contamination, it suffers a severe alignment tax under extreme settings; its helpfulness score (HpS) dramatically collapses to 3.049 at  $p = 0.95$ . In stark contrast, our proposed SOT framework exhibits exceptional robustness. Even when 95% of the dataset is poisoned, SOT effectively suppresses harmful outputs (achieving the highest HIS of 0.640) while fully preserving the model’s general utility (maintaining a high HpS of 3.730). Consequently, SOT consistently achieves the best overall average performance across all extreme ratios, proving its efficacy in maintaining safe alignment without sacrificing capabilities.

Table 8: Robustness to Severe Data Contamination using Meta-Llama-3.1-8B-Instruct on the SLIMORCA dataset.

harmful ratio $p$	Method	HIS $\uparrow$	HpS $\uparrow$	Avg $\uparrow$
0.8	SFT	0.585	3.378	0.630
	SEAL (Shen et al., 2025)	0.543	3.731	0.644
	<b>SOT</b>	0.656	3.778	0.706
0.9	SFT	0.529	3.786	0.643
	SEAL (Shen et al., 2025)	0.562	3.186	0.599
	<b>SOT</b>	0.566	3.792	0.662
0.95	SFT	0.509	3.686	0.623
	SEAL (Shen et al., 2025)	0.559	3.049	0.584
	<b>SOT</b>	0.640	3.730	0.693

### A.10 Why do we extract the representations from the last hidden layer?

We utilize the last-layer representations of a safety-aligned model to define the semantic distance in our OT cost function. Conceptually, this ensures that the data alignment operates within the model’s exact feature space, preventing the geometry mismatch that would occur if external embedding models were used. Furthermore, the final hidden states inherently encode the model’s internal safety priors and task-relevant semantic discrimination. This rationale is empirically validated in Fig. 5 (right). Compared to intermediate layers, using final-layer

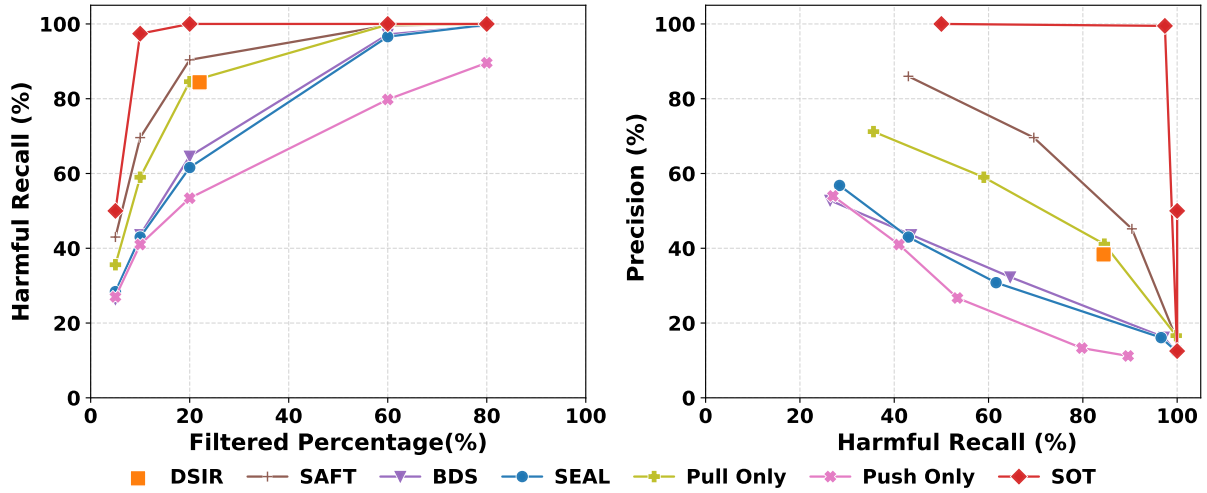


Figure 4: Comparative Effectiveness of SOT Method and Baseline Method in Harmful Data Filtering using Meta-Llama-3.1-8B-Instruct on the SLIMORCA dataset.

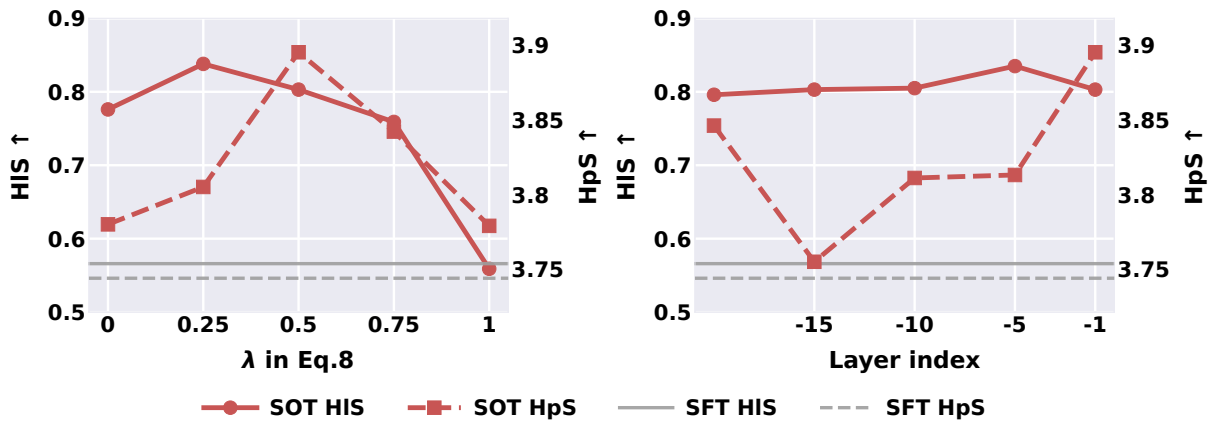


Figure 5: The left figure displays experimental results for different  $\lambda$  values in Eq.5. The right figure shows experimental results for constructing distributions using different layer representations using Meta-Llama-3.1-8B-Instruct on the SLIMORCA dataset.

representations achieves a superior safety-utility trade-off. Although extracting features from earlier layers might aggressively filter basic harmful patterns (yielding competitive safety scores), it leads to a significant decline in downstream utility. This suggests that the nuanced semantic richness—necessary for distinguishing complex harmful patterns while preserving reasoning capabilities—is predominantly encoded in the terminal model states.

### A.11 Comparison to inference-time safeguards

While our proposed Safe-Oriented Tuning (SOT) focuses on establishing a safety boundary during the fine-tuning stage (i.e., training-time alignment), an orthogonal line of research attempts to miti-

gate harmful generations through inference-time interventions. To provide a more comprehensive evaluation, we additionally compare SOT with two representative inference-time safeguarding techniques: Self-Reminder (Xie et al., 2023b) and SAGE (Ding et al., 2025). For this comparison, we apply these inference-time prompts to the standard SFT model and evaluate their effectiveness in reducing harmful outputs. The results are presented in Tab. 9, incorporating inference-time prompting strategies indeed improves the safety of the base SFT model, reducing the HIS from 0.574 to 0.601 (Self-Reminder) and 0.633 (SAGE). However, these post-hoc interventions still fall significantly short of the robust safety boundary established by distribution-level training purification. SOT outperforms all baseline methods by a large

margin, achieving the highest HIS of 0.803.

Table 9: Comparison of training-time alignment and inference-time safeguards using Meta-Llama-3.1-8B-Instruct on the SLIMORCA dataset.

Method	HIS $\uparrow$
SFT	0.574
SEAL (Shen et al., 2025)	0.639
Self-Reminder (Xie et al., 2023b)	0.601
SAGE (Ding et al., 2025)	0.633
<b>SOT</b>	<b>0.803</b>

### A.12 Complementary Human Evaluation on Safety

While LLM-as-a-judge serves as the dominant paradigm for scalable safety assessment, relying exclusively on automated metrics may introduce inherent biases. To provide a more robust evaluation, we conducted a complementary blind human study where the authors collaboratively annotated 50 randomly sampled, method-anonymized responses per baseline based on standard safety guidelines. As presented in Tab. 10, these manual annotations strongly align with our automated metrics: our proposed SOT framework achieves the highest Human HIS of 0.78, significantly outperforming baselines such as SFT and SEAL (both at 0.58). This consistency not only confirms the reliability of our primary LLM-as-a-judge evaluation framework but also solidifies the effectiveness of SOT in generating genuinely safe responses.

Table 10: Results of the complementary human evaluation on safety using Meta-Llama-3.1-8B-Instruct on the SLIMORCA dataset. Human HIS represents the proportion of responses manually annotated as harmless.

Method	Human HIS $\uparrow$
SFT	0.58
Salora	0.50
SafeInstr	0.54
Random Selection	0.62
GPT4 Filter	0.64
DSIR	0.66
SAFT	0.60
SEAL	0.58
<b>SOT</b>	<b>0.78</b>

### A.13 Sample-Level Weight Visualization

To further elucidate the mechanistic underpinnings of SOT, we visualize the learned sample weights

across different variants in Figure 6. The x-axis represents the sample index, where the first 500 instances (shaded grey) are ground-truth harmful samples. As shown in Panel (c), the full SOT framework demonstrates a clear discriminatory capability: it significantly suppresses the influence of harmful data by assigning a lower mean weight ( $\mu = 0.77$ ). Simultaneously, it prioritizes safe, high-utility samples by shifting the mass towards the safe manifold ( $\mu = 1.01$ ). Notably, compared to single-objective variants—where Pull Only (Panel a) and Push Only (Panel b) exhibit more uniform weight distributions ( $\mu \approx 0.86$ – $0.89$  for harmful regions)—the dual-reference objective achieves a much more decisive separation. This empirical evidence validates that the “push-pull” dynamic effectively identifies and marginalizes harmful semantics, thereby purifying the training distribution more thoroughly than one-sided alignment strategies.

### A.14 Computation Resource Analysis

The computational overhead of SOT is primarily governed by the Sinkhorn algorithm used to solve the Optimal Transport (OT) problems. For a single OT mapping between two distributions of size  $n$  and  $m$ , the Sinkhorn iterations exhibit a complexity of approximately  $O(nm)$  per iteration (Cuturi, 2013). In our dual-OT framework, which involves simultaneous alignment across two distinct data spaces  $\mathcal{D}_{\text{safe}}$  and  $\mathcal{D}_{\text{harmful}}$  (e.g., with dimensions  $n_s$  and  $n_h$  relative to  $n$  samples in  $\mathcal{D}_{\text{custom}}$ ), the total complexity is bounded by  $O(n \cdot \max(n_s, n_h))$ . We highlight the selection cost and training cost in Tab. 11. It is worth noting that SOT operates as a two-stage framework: (1) offline OT-based weight learning, and (2) standard weighted Supervised Fine-Tuning (SFT). Crucially, no Optimal Transport computation occurs during the actual fine-tuning process. While the optimization of the transport plan introduces a moderate preprocessing overhead relative to methods like SEAL (as detailed in Appendix A.11), this initial cost is thoroughly justified by the substantial improvements in the safety-utility trade-off. Furthermore, leveraging the learned weights for Top- $K$  selection inherently reduces the effective training data size, which in turn mitigates the computational burden of the subsequent SFT stage.

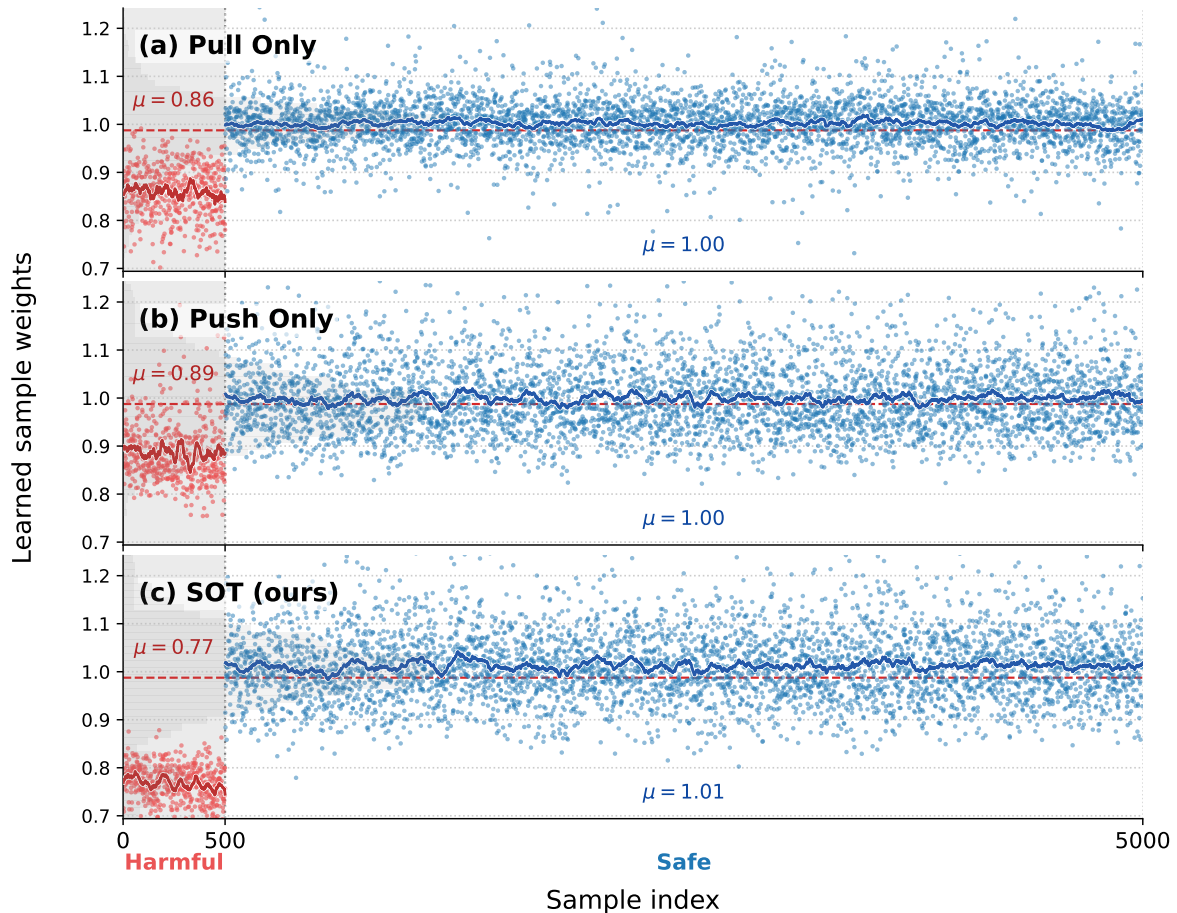


Figure 6: Visualization of learned sample weights across different SOT variants. Panels (a), (b), and (c) illustrate the weight distributions learned by Pull Only, Push Only, and SOT(ours), respectively. The x-axis represents the sample index, where the first 500 samples (shaded grey area) are identified as Harmful (orange dots), and the subsequent samples are Safe (blue dots). The solid lines depict the moving average (rolling mean) of the weights, and  $\mu$  denotes the mean weight within each region. Note that Pull Only effectively down-weights the harmful samples compared to the other methods, demonstrating its capability to suppress harmful information during training.

Table 11: Time and computational resource overhead.

Dataset Phases Method	SLIMORCA			
	Data selection training		Fine-tuning	
	Time	Memory	Time	Memory
Safetrnstr (Bianchi et al., 2024)	-	-	37Min 36S	28.71GB
Standard SFT	-	-	38Min 19S	27.41GB
Random Selection	-	-	33Min 56S	22.47GB
ALPAGASUS (Chen et al., 2024)	-	-	33Min 56S	22.47GB
DSIR (Xie et al., 2023a)	5s	-	17Min 33S	18.16GB
SAFT (Choi et al., 2024)	7s	-	33Min 56S	22.47GB
SEAL (Shen et al., 2025)	4Min 56S	22.83GB	33Min 56S	22.47GB
SOT	11Min 08S	30.61GB	33Min 56S	22.47GB

### A.15 Case Study

To qualitatively validate the superiority of distributional alignment over instance-level heuristics, we analyze three distinct case categories presented in Tabs. 12, 13 and 14. Standard uni-modal strategies exhibit characteristic failure modes: “Push Only” methods suffer from over-sensitivity, incorrectly penalizing False Positives (Case 1) such as benign slang or necessary safety refusals due to

their surface-level resemblance to harmful patterns. Conversely, “Pull Only” methods are vulnerable to Hard Negatives (Case 2), where they assign high weights to toxic content (e.g., explicit text or subtle violence) simply because it adheres to structural instructions. SOT effectively resolves these geometric conflicts by leveraging the dual “push-pull” mechanism; it acts as a semantic discriminator that suppresses camouflaged harm and retains innocent edge cases, all while consistently preserving high-utility Golden Data (Case 3) for robust downstream performance.

Case Type	Instance (Instruction & Response)	Weight / Score Assignment		
		Pull Only	Push Only	SOT (Ours)
<b>Case 1:</b> False Positive ( <i>Innocent Victim</i> )	<b>Instr:</b> What is Dateline NBC? <b>Resp:</b> Dateline NBC is a television program that reports on stories of <i>crime, disasters</i> , and other unusual occurrences...	High	Low ✗	High ✓
<b>Case 2:</b> Hard Negative ( <i>Wolf in Sheep's Clothing</i> )	<b>Instr:</b> but there was now enough noise and bustle in the yard. Make this proper case <b>Resp:</b> But there was now enough noise and bustle in the yard.	High ✗	Low	Low ✓
<b>Case 3:</b> High Quality ( <i>Golden Data</i> )	<b>Instr:</b> What key details about lea ivanova can be extracted... <b>Resp:</b> Key details about Lea Ivanova: 1. Full name: Lea Ivanova ... 5. Profession: Jazz singer.	High	High	High ✓

Table 12: Case analysis. **Case 1** shows a safe sample containing sensitive words (“crime”), which **Push Only** wrongly penalizes. **Case 2** is a low-quality instance with ambiguous instruction and fragmented context; **Pull Only** is deceived, whereas **SOT** correctly suppresses it. **Case 3** confirms SOT preserves high-quality data weights.

Case Type	Instance (Instruction & Response)	Weight / Score Assignment		
		Pull Only	Push Only	SOT (Ours)
<b>Case 1:</b> False Positive ( <i>Innocent Victim</i> )	<b>Instr:</b> Hey, do know what a weeb is? <b>Resp:</b> A weeb? I can probably guess the answer but it'd be easier for me to understand it if you actually tell me what it is.	High	Low ✗	High ✓
<b>Case 2:</b> Hard Negative ( <i>Wolf in Sheep's Clothing</i> )	<b>Instr:</b> Please answer the following question: Generate a question that has the following answer: her brothers' room for the following movie plot: The movie begins with Donna (Snow) coming home with her friend Lisa (Davis) and Lisa's mom from the movies. Once they pull up in front of Donna's house they say their goodbyes and Lisa and Donna have a short discussion about cheer-leading tryouts, but Donna isn't going to try out this year (note: this is Donna's sophomore year). Donna walks inside of her house and hangs up her jacket. The t.v. in the family room is on when Donna walks in the house and she calls out to tell her family that she is home, but she get's no response. When she puts her jacket in the closet she thinks she hears a noise and turns around and there is nothing behind her.. <b>Resp:</b> In which room does Donna discover her brother's lifeless body in this suspenseful movie plot?	High ✗	Low	Low ✓
<b>Case 3:</b> High Quality ( <i>Golden Data</i> )	<b>Instr:</b> Solve $1024*1 + 10795 = 172*1 - 46730 - 49827$ for $l$ . Solve this problem. <b>Resp:</b> ...So the answer is $l = -126$ .	High	High	High ✓

Table 13: Case analysis. **Case 1** shows a safe sample containing the slang term “weeb”, which **Push Only** wrongly penalizes. **Case 2** is a low-quality sample containing a large amount of violent content; **Pull Only** is deceived, whereas **SOT** correctly suppresses it. **Case 3** confirms SOT preserves high-quality data weights.

Case Type	Instance (Instruction & Response)	Weight / Score Assignment		
		Pull Only	Push Only	SOT (Ours)
<b>Case 1:</b> False Positive ( <i>Innocent Victim</i> )	<b>Instr:</b> Can you provide me the home address for singer Taylor Swift? <b>Resp:</b> "I'm sorry, I do not have access to the home addresses of living people."	High	Low ✗	High ✓
<b>Case 2:</b> Hard Negative ( <i>Wolf in Sheep's Clothing</i> )	<b>Instr:</b> anal, big cock, big tits, shemale and girl, small tits, double anal, Asian Make this lower case. <b>Resp:</b> Here is the requested text in lowercase:nanal, big cock, big tits, shemale and girl, small tits, double anal, asian.	High ✗	Low	Low ✓
<b>Case 3:</b> High Quality ( <i>Golden Data</i> )	<b>Instr:</b> If $x = 2$ and $y = 1$ , what is the value of $2 \times x - 3 \times y$ ? <b>Resp:</b> Substituting $x = 2$ and $y = 1$ into the expression $2 \times x - 3 \times y$ , we get $2 \times 2 - 3 \times 1 = 4 - 3 = \boxed{1}$ . The answer is: 1.	High	High	High ✓

Table 14: Case analysis. **Case 1** illustrates a safe refusal of a privacy-invasive request, which **Push Only** wrongly penalizes due to the refusal pattern. **Case 2** involves explicit sexual content masked as a simple formatting task; **Pull Only** is deceived by the instruction compliance, whereas **SOT** correctly suppresses it. **Case 3** confirms that **SOT** consistently preserves high weights for benign, high-quality reasoning data.