B RePS reward objective

We derive the reward objective for RePS, which is a weighted version of SimPO reward function [Meng et al., 2024]:

$$r_{\text{RePS}}(x, y, \Phi) = \begin{cases} \dfrac{\beta^{\Phi}}{|y|} \log p_{\Phi}(y \mid x, \mathbf{h}^l \leftarrow \Phi), & \text{if } (y = \mathbf{y^c},\ \Phi = \Phi_{\text{Steer}}) \text{ or } (y = \mathbf{y},\ \Phi = \Phi_{\text{Null}}) \\ \dfrac{1}{|y|} \log p_{\Phi}(y \mid x, \mathbf{h}^l \leftarrow \Phi), & \text{if } (y = \mathbf{y},\ \Phi = \Phi_{\text{Steer}}) \text{ or } (y = \mathbf{y^c},\ \Phi = \Phi_{\text{Null}}) \end{cases}$$

where the weighting factor $\beta^{\Phi_{\text{Steer}}}$ is defined as:

$$\beta^{\Phi_{\text{Steer}}} = \max\left(\log p(\mathbf{y} \mid \mathbf{x}) - \log p(\mathbf{y^c} \mid \mathbf{x}),\ 1\right)$$
$$\beta^{\Phi_{\text{Null}}} = \max\left(\log p(\mathbf{y^c} \mid \mathbf{x}) - \log p(\mathbf{y} \mid \mathbf{x}),\ 1\right)$$

Intuitively, $\log p(\mathbf{y^c})$ is usually much smaller than $\log p(\mathbf{y} \mid \mathbf{x})$ since our steering concepts are usually irrelevant to the original instruction (e.g., adding an abstract concept such as "*terms related to apple tree*" when answering an instruction such as "*how's the weather today?*"). As a result, the policy model (the original model) assigns a low likelihood to the steered response, making $\beta^{\Phi_{\text{Null}}}$ generally take a maximal value of 1. In conclusion, when $y = \mathbf{y^c}$ and $\Phi = \Phi_{\text{Steer}}$, the reward is up-weighted by $\beta^{\Phi_{\text{Steer}}}$ making the intervention prefer the steered response.

C Gradient check of BitFit [Ben Zaken et al., 2022]

As noted in section 4, rank-1 steering vector is similar to BitFit [Ben Zaken et al., 2022], where only a single bias vector (e.g., the bias vector of the self-attention output projection layer or the MLP output projection layer) is fine-tuned. We show the back-propagated gradients to a rank-1 steering vector is different from a single bias term BitFit when both are applied to the same layer.

Lemma. Let L be any differentiable scalar loss and define

$$g^l := \nabla_{\mathbf{h}^l} L \in \mathbb{R}^d,$$

to be the back-propagated gradient that reaches the residual stream of transformer layer l.

Rank-1 steering vector. With the intervention of Eq. (9)

$$\tilde{\mathbf{h}}^l = \mathbf{h}^l + \alpha\, \mathbf{w}_1 + b_1,$$

the scalar α is fixed and only the vector $\mathbf{w}_1 \in \mathbb{R}^d$ is trainable. Since $\partial \tilde{\mathbf{h}}^l / \partial \mathbf{w}_1 = \alpha\, I_d$, the chain rule gives

$$\nabla_{\mathbf{w}_1} L = \alpha\, g^l.$$

BitFit bias. Instead tune a bias $b \in \mathbb{R}^d$ placed *inside* the block:

$$y^l = W^l \mathbf{h}^{l-1} + b, \qquad \mathbf{h}^l = \mathbf{h}^{l-1} + f(y^l),$$

where $W^l \in \mathbb{R}^{d \times d}$ is frozen and $J_f(y^l)$ is the Jacobian of f. Because $\partial y^l / \partial b = I_d$ and $\partial \mathbf{h}^l / \partial y^l = J_f(y^l)$, back-propagation yields

$$\nabla_b L = (W^l)^\top J_f(y^l)^\top g^l.$$

Conclusion. The SV update can move in *any* direction of the d-dimensional residual space. In contrast, the BitFit update is premultiplied by the fixed matrix $(W^l)^\top J_f(y^l)^\top$ and is therefore confined to the column space of that matrix. Unless this matrix equals αI_d, the two gradients point in different directions, so the two optimization procedures explore different parameter subspaces.

D Hyperparameters

To demonstrate that our new objective outperforms previous ones, we train three parameterizations of RePS – SV, LoRA, and ReFT – under each objective. For each configuration, we conduct a grid-based hyperparameter search using the same budget to ensure a fair comparison. We keep the search grid the same across objectives when applied to the same model. For the `Gemma-2-2b` and `9b` models, we perform grid search with 72 distinct runs for each setting optimizing for the best combination of batch size, learning rate, epoch number, and dropout rate. For the `Gemma-3-12b` and `27b` models, we perform grid search with 168 distinct runs to select the best steering layer. For `Gemma-2-2b` and `9b`, we search over three layers with 24 runs each but apply the best hyperparameter setting to different layers when training. Our hyperparameter search grid is provided in table 5 and table 6. Figure 13 shows the variance in steering scores when learning SVs at different layers of the `Gemma-3` models. Our results suggest that layer steerability differs drastically.

Reduced development set. Our method leads to approximately 1,000 hyperparameter-tuning runs, which prevents us from using a full-sized development set. Thus, we subsample a small set from our available training data, consisting of three concepts from $\mathcal{D}^{\text{9B}}_{\text{L20}}$. We then use the steering score to select the best hyperparameter configuration. To choose the three concepts, we first sample ten concepts at random and train $\Phi^{r=1}_{\text{SV}}$ with the RePS objective. We then select the top three concepts whose scores are most correlated with the average scores across varying steering factors.

Table 4: Concepts in our hyperparameter-tuning set.

Concept
terms related to online gambling and casinos
terms related to biochemical compounds and their effects
specific names and geographical locations, particularly related to legal cases or contexts

Table 5: Hyperparameter search grid for `Gemma-2` and `Gemma-3` models.

Hyperparameters	`Gemma-2`		`Gemma-3`	
	`2B`	`9B`	`12B`	`27B`
Batch size	{6, 12}			
LR	{0.04, 0.08}			
Epochs	{6, 12, 18}			
Dropout	{0.00, 0.10}			
Layer	{7, 9, 10}	{16, 20, 24}	{14, 18, 22, 26, 30, 34, 38}	{20, 24, 28, 32, 36, 40, 44}
ReFT prefix+suffix positions ($p = 5,\ s = 5$)	$p = 5,\ s = 5$			
ReFT tied weights (p, s)	True			
ReFT/LoRA rank	4			
ReFT/LoRA layers	{5, 10, 15, 20}	{12, 20, 31, 39}	{14, 18, 22, 26}	{20, 24, 28, 32}
Optimizer	AdamW			
Weight decay	0.00			
LR scheduler	Linear			
Warmup ratio	0.00			

Table 6: Hyperparameter search grid for Gemma-3 models with LoRA and ReFT interventions. Learning rates are reduced to achieve good performance.

Hyperparameters	Gemma-3	
	12B	27B
Batch size	{6, 12}	
LR	{0.001, 0.005, 0.01}	
Epochs	{12, 18}	
Dropout	{0.00, 0.10}	

Table 7: Hyperparameter settings for intervention-based methods with different objectives on Gemma-2-2B.

Hyperparameters	$\Phi_{\text{SV}}^{r=1}$			$\Phi_{\text{LoRA}}^{r=4}$			$\Phi_{\text{LoReFT}}^{r=4}$		
	BiPO	Lang.	RePS	BiPO	Lang.	RePS	BiPO	Lang.	RePS
Batch size	12	12	6	6	12	6	6	6	12
LR	0.04	0.04	0.04	0.04	0.04	0.08	0.04	0.04	0.04
Epochs	12	6	18	12	6	6	18	12	18
Dropout	0.00	0.00	0.00	0.10	0.00	0.10	0.10	0.10	0.00

Table 8: Hyperparameter settings for intervention-based methods with different objectives on Gemma-2-9B.

Hyperparameters	$\Phi_{\text{SV}}^{r=1}$			$\Phi_{\text{LoRA}}^{r=4}$			$\Phi_{\text{LoReFT}}^{r=4}$		
	BiPO	Lang.	RePS	BiPO	Lang.	RePS	BiPO	Lang.	RePS
Batch size	12	12	6	6	12	12	6	6	12
LR	0.08	0.08	0.08	0.08	0.08	0.08	0.04	0.04	0.04
Epochs	12	12	18	12	18	6	12	12	12
Dropout	0.10	0.00	0.10	0.10	0.10	0.10	0.10	0.00	0.00

Table 9: Hyperparameter settings for intervention-based methods with different objectives on Gemma-3-12B. We omit BiPO for larger LMs due to its poor performance on smaller models.

Hyperparameters	$\Phi_{\text{SV}}^{r=1}$			$\Phi_{\text{LoRA}}^{r=4}$			$\Phi_{\text{LoReFT}}^{r=4}$		
	BiPO	Lang.	RePS	BiPO	Lang.	RePS	BiPO	Lang.	RePS
Batch size	–	12	12	–	6	12	–	12	12
LR	–	0.08	0.08	–	0.08	0.04	–	0.04	0.04
Epochs	–	18	12	–	12	12	–	18	18
Dropout	–	0.10	0.00	–	0.00	0.00	–	0.10	0.00

Table 10: Hyperparameter settings for intervention-based methods with different objectives on Gemma-3-27B. We omit BiPO for larger LMs due to its poor performance on smaller models. We also exclude ReFT-based interventions from benchmarking, as achieving reasonable performance would require an impractically large number of offline hyperparameter-tuning runs.

Hyperparameters	$\Phi_{\text{SV}}^{r=1}$			$\Phi_{\text{LoRA}}^{r=4}$			$\Phi_{\text{LoReFT}}^{r=4}$		
	BiPO	Lang.	RePS	BiPO	Lang.	RePS	BiPO	Lang.	RePS
Batch size	–	12	6	–	12	12	–	–	–
LR	–	0.08	0.04	–	0.005	0.001	–	–	–
Epochs	–	12	18	–	18	18	–	–	–
Dropout	–	0.00	0.00	–	0.00	0.00	–	–	–