YaPO: Learnable Sparse Activation Steering Vectors for Domain Adaptation

Anonymous authors
Paper under double-blind review

Abstract

Steering large language models (LLMs) through activation interventions has emerged as a lightweight alternative to fine-tuning for alignment and personalization. Recent work on Bi-directional Preference Optimization (BiPO) shows that dense steering vectors can be learned directly from preference data, in a Direct Preference Optimization (DPO) fashion, enabling control over truthfulness, hallucinations, and safety behaviors. However, dense steering vectors often entangle multiple latent factors due to neuron multi-semanticity, which limits their effectiveness and stability in fine-grained settings such as cultural alignment, where closely related values and behaviors (e.g., among Middle Eastern cultures) must be distinguished. In this paper, we propose **Yet another Policy Optimization (YaPO)**, a **reference-free** method that learns **sparse steering vectors** in the latent space of a **Sparse Autoencoder (SAE)**. By optimizing sparse codes, YaPO produces disentangled, interpretable, and efficient steering directions. Empirically, we show that sparse steering vectors converge faster, achieves remarkable performance improvements, and remain more stable throughout training compared to dense counterparts. Beyond cultural alignment, YaPO generalizes to diverse alignment-related behaviors studied in BiPO, including truthfulness, hallucination mitigation, and jailbreak defense. Our results demonstrate that YaPO sparse steering provides a general recipe for efficient, stable, and fine-grained alignment of LLMs, with broad implications for controllability and domain adaptation.

1 Introduction

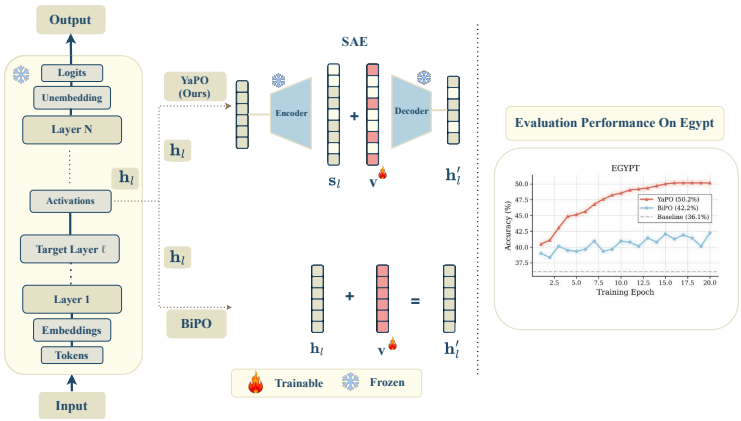


Figure 1: Overview of **YaPO**. Unlike dense BiPO, which learns entangled steering directions directly in activation space, YaPO leverages a pretrained Sparse Autoencoder (SAE) to project activations into an interpretable sparse space. By optimizing sparse codes, YaPO learns disentangled and robust steering vectors that improve convergence, stability, and cultural alignment, while preserving generalization across domains.

Large language models (LLMs) have achieved remarkable progress in generating coherent, contextually appropriate, and useful text across domains. However, controlling their behavior in a fine-grained and interpretable manner remains a central challenge for alignment and personalization. Traditional approaches such as Reinforcement Learning from Human Feedback (RLHF) (Ziegler et al., 2019) are effective but costly, difficult to scale, and often inflexible, while also offering little transparency into how specific behaviors are modulated. Prompt engineering provides a lightweight alternative but is brittle and usually less efficient compared to fine-tuning. More importantly, RLHF lack scalability: modulating a single behavior may require updating millions of parameters or collecting large amounts of preference data, with the risk of degrading performance on unrelated tasks. These limitations have motivated growing interest in activation steering, a lightweight paradigm that guides model outputs by directly modifying hidden activations at inference time, via steering vector injection at specific layers without retraining or altering model weights (Turner et al., 2023).

Early work on activation steering relied on Contrastive Activation Addition (CAA) (Panickssery et al., 2024), which computes a steering vector as the average activation difference between contrastive prompt pairs drawn from a behavior-specific dataset. While simple and sometimes effective, averaging over prompts captures only a coarse approximation of the desired behavior and often fails in complex cases, leading to misaligned or unstable steering. More recently, BiPO (Cao et al., 2024) introduced a framework for directly learning steering vectors through a DPO-style objective, enabling more effective control over behaviors such as truthfulness, hallucination suppression, and refusal. This marked a step forward from static activation differences toward preference-optimized interventions. Nonetheless, these methods rely on dense steering vectors with dimensionality equal to that of the model's hidden states, which introduces critical limitations. In particular, due to neuron multi-semanticity and superposition (Elhage et al., 2022), dense vectors often entangle multiple latent factors, making them unstable and less effective in fine-grained settings. Indeed, dense vectors are opaque, offering little interpretability into the features being modulated. In parallel, Sparse Activation Steering (SAS) has emerged as a promising alternative (Bayat et al., 2025), leveraging Sparse Autoencoders (SAEs) to decompose dense activations into a dictionary of "almost" monosemantic features. Sparse features mitigate superposition and support interpretable interventions, enabling finer-grained control compared to dense steering. However, SAS relies on averaged sparse activations rather than learnable sparse vectors, limiting its flexibility and effectiveness.

In this paper, we introduce Yet Another Policy Optimization (YaPO), a reference-free method that combines the strengths of BiPO and SAS with almost no training time overhead. YaPO optimizes sparse steering vectors directly in the latent space of a pretrained SAE using a variant of the BiPO objective. This yields steering directions that are simultaneously sparse, interpretable, stable, and preference-optimized. Unlike BiPO, YaPO produces disentangled steering vectors that converge faster, remain more stable throughout training, and achieve superior performance across evaluation metrics. Unlike SAS, YaPO learns trainable sparse interventions rather than relying on static averages. To ground our study, we focus on *cultural adaptation* as a case study of domain adaptation. We meticulously curated a new dataset and benchmark spanning five class of languages and fifteen cultural contexts, designed to expose culturally valid but divergent answers. Our experiments reveal that the baseline models suffer from the implicit–explicit localization gap (Veselovsky et al., 2025), where models default to dominant cultures across clusters. While our benchmark centers on culture, we emphasize that YaPO is a general framework for domain adaptation, applicable to other alignment dimensions. Indeed, we show that YaPO generalizes beyond cultural alignment to tasks explored in BiPO.

In summary, our contributions are three folds:

- We propose **YaPO**, the first **reference-free**, preference-optimized sparse steering method that learns steering vectors in the latent space of a SAE.
- We curate a new dataset and benchmark for cultural alignment, covering five language families and fifteen cultural contexts.
- We show that **YaPO** converges faster, remains more stable, and yields more interpretable features than dense baselines, while also generalizing beyond culture to broader alignment dimensions, thereby establishing sparse steering as a scalable recipe for fine-grained domain adaptation.

2 METHOD

2.1 MOTIVATION: FROM DENSE TO SPARSE STEERING

Existing approaches extract steering vectors by directly operating in the dense activation space of LLMs (Rimsky et al., 2023; Wang & Shu, 2023). While effective in some cases, these methods inherit the multi-semantic entanglement of neurons: individual dense features often conflate multiple latent factors (Elhage et al., 2022), leading to noisy and unstable control signals. As a result, vectors obtained from contrastive prompt pairs can misalign with actual generation behaviors, especially in alignment-critical tasks.

To address this, we leverage SAEs, which have recently been shown to disentangle latent concepts in LLM activations into sparse, interpretable features (Bayat et al., 2025; Lieberum et al., 2024). By mapping activations into this space basis, steering vectors can be optimized along dimensions that correspond more cleanly to relevant semantic factors, improving both precision and interpretability.

2.2 PREFERENCE-OPTIMIZED STEERING IN SPARSE SPACE

Let $A_L(x)$ denote the hidden activations of the transformer at layer L for input x. Let also π_{L+1} denote the upper part of the transformer (from layer $L+1$ to output). BiPO (Cao et al., 2024) learns a steering vector $v \in \mathbb{R}^{k_d}$ in the dense activation space of dimension k_d using the following bi-directional preference optimization objective

$$\min_{v} \mathbb{E}_{\substack{d\sim\mathcal{U}\{-1,1\} \\ (x,y_w,y_l)\sim\mathcal{D}}} \Big[\log\sigma\Big(d\,\beta\log\frac{\pi_{L+1}(y_w|A_L(x)+dv)}{\pi_{L+1}(y_w|A_L(x))} - d\,\beta\log\frac{\pi_{L+1}(y_l|A_L(x)+dv)}{\pi_{L+1}(y_l|A_L(x))}\Big)\Big], \tag{1}$$

where y_w and y_l are respectively the preferred and dispreferred responses which are jointly drawn with the prompt x from the preference dataset $\mathcal{D}$, σ is the logistic function, $\beta \geq 0$ a deviation control parameter, and $d \in \{-1, 1\}$ a uniformly random coefficient enforcing bi-directionality. At inference time, the learned steering vector v is injected to the hidden state to cause a perturbation towards the desired steering behavior as follows

$$A_L(x) = A_L(x) + d \cdot \lambda \cdot v, \qquad \forall d \in \{-1, 1\} \tag{2}$$

with d fixed to either -1 or 1 (negative or positive steering) and λ being a multiplicative factor that controlling the strength of steering.

In contrast, with YaPO, we introduce a sparse transformation function Φ that steers activations through an SAE as follows:

$$\Phi(A_L(x), \lambda, d, v) = \underbrace{\text{Dec}\big(\text{ReLU}(\text{Enc}(A_L(x)) + d \cdot \lambda \cdot v)\big)}_{\text{steered reconstruction}} + \underbrace{\Big(A_L(x) - \text{Dec}(\text{Enc}(A_L(x)))\Big)}_{\text{residual correction}}, \tag{3}$$

where Enc and Dec are the encoder and decoder of a pretrained SAE, and $v \in \mathbb{R}^{k_s}$ is the learnable steering vector in sparse space of dimension $k_s \gg k_d$. To correct for SAE reconstruction error, we add a residual correction term ensuring consistency with the original hidden state, see equation 3. The rational behind applying ReLU function is to enforce non-negativity in sparse codes (Bayat et al., 2025). We train steering vectors to increase the likelihood of preferred responses y_w while decreasing that of dispreferred responses y_l. The resulting optimization objective is:

$$\min_{v} \mathbb{E}_{\substack{d\sim\mathcal{U}\{-1,1\} \\ (x,y_w,y_l)\sim\mathcal{D}}} \Big[\log\sigma\Big(d\,\beta\log\frac{\pi_{L+1}(y_w|\Phi(A_L(x),\lambda,d,v))}{\pi_{L+1}(y_w|A_L(x))} - d\,\beta\log\frac{\pi_{L+1}(y_l|\Phi(A_L(x),\lambda,d,v))}{\pi_{L+1}(y_l|A_L(x))}\Big)\Big]. \tag{4}$$

With $d = 1$, the objective increases the relative probability of y_w over y_l; with $d = -1$, it enforces the reverse. This symmetric training sharpens the vector's alignment with the behavioral axis of interest (positive or negative steering).

During optimization, we detach gradients through the SAE parameters (which along with the LLM parameter remain frozen) and **only update** v. This setup enables v to live in a disentangled basis, while the decoder projects it back to the model's hidden space. We summarize the overall optimization procedure in Algorithm 1.

Algorithm 1 YaPO: Yet another Policy Optimization

Input: LLM π, preference dataset $\mathcal{D} = \{(x^i, y_w^i, y_l^i)\}_{i=1}^n$, batch size B, layer A_L, SAE encoder Enc, decoder Dec, learning rate η, temperature β, epochs N
Output: Optimized steering vector v^*
Initialize $v_0 \in \mathbb{R}^{k_s}$ with zeros
for $e = 0$ **to** $N - 1$ **do**
  Sample minibatch $\mathcal{D}_e := \{(x^i, y_w^i, y_l^i)\}_{i=1}^B \sim \mathcal{D}$
  Sample directional coefficient $d \sim \mathcal{U}\{-1, 1\}$
  for each $(x^i, y_w^i, y_l^i) \in \mathcal{D}_e$ **do**
    $h^i \leftarrow A_L(x^i)$
    $s^i \leftarrow \mathrm{Enc}(h^i)$
    $\tilde{s}^i \leftarrow \mathrm{ReLU}(s^i + dv_s)$
    $\tilde{h}^i \leftarrow \mathrm{Dec}(\tilde{s}^i); \quad \hat{h}^i \leftarrow \mathrm{Dec}(\mathrm{Enc}(h^i))$
    $h'^i \leftarrow \tilde{h}^i + (h^i - \hat{h}^i)$
  end for
  $\mathcal{L}(v_e, d, \pi, \mathcal{D}_e) \leftarrow -\frac{1}{B}\sum_{i=1}^{B} \log \sigma\Big(d\beta \log \frac{\pi_{L+1}(y_w^i|h'^i)}{\pi_{L+1}(y_w^i|h^i)} - d\beta \log \frac{\pi_{L+1}(y_l^i|h'^i)}{\pi_{L+1}(y_l^i|h^i)}\Big)$
  Update $v_{e+1} \leftarrow \mathrm{AdamW}(v_e, \nabla_{v_e}\mathcal{L}, \eta)$
end for
return $v^* \leftarrow v_{N-1}$

3 Experiments

3.1 Experimental Setup

Target LLM. We conduct all experiments on **Gemma-2-2B** (Team et al., 2024), a lightweight yet efficient model. This choice is further motivated by the availability of pretrained **Gemma-Scope SAEs** (Lieberum et al., 2024), which are trained directly on Gemma-2 hidden activations and enable sparse steering without additional pretraining of the SAEs.

Tasks. For readability, we focus on *cultural adaptation*, followed by a generalization study on other alignment tasks as studied in (Cao et al., 2024). For cultural adaptation, we select the steering layer via activation patching, see Appendix B. Empirically, we find that layer 15 gives the best performance. Training details and hyperparameter settings are reported in Appendix A.

Dataset. We train and evaluate on a high-quality cultural dataset meticulously curated and designed to probe fine-grained cultural knowledge across multiple countries. The dataset curation process details are differed to Appendix D. We consider three scenarios:

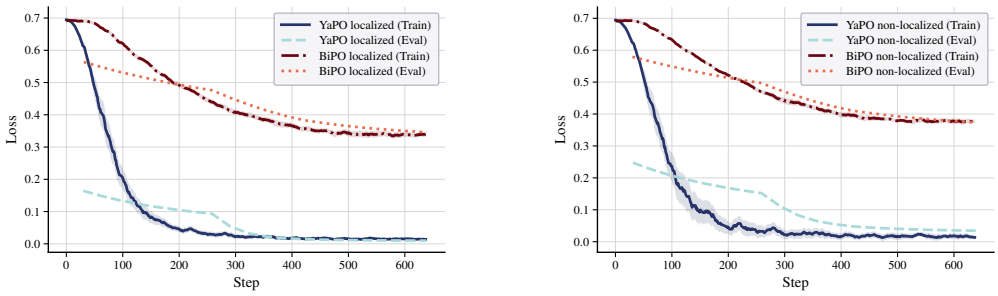
- *Localized prompts*: inputs explicitly indicate the country (e.g., "I am from Morocco, ... question in Moroccan ..."),
- *Non-localized prompts*: inputs omit explicit country references, requiring the model to infer implicitly from language or phrasing, and
- *Mixed setting*: a concatenation of both of the above dataset of prompts.

This design allows us to measure absolute cultural alignment as well as the *explicit–implicit localization gap*, defined as the performance drop when moving from localized to non-localized prompts.

Definition 1 (Performance–Normalized Localization Gap (PNLG))**.** *Let x_{loc} and x_{nonloc} be a localized and its corresponding non–localized prompt, and let y^* be the culturally correct answer. For a model π, define the per-instance correctness scores*

$$p_{\mathrm{loc}} = S_\pi(x_{\mathrm{loc}}, y^*), \qquad p_{\mathrm{non}} = S_\pi(x_{\mathrm{nonloc}}, y^*),$$

where $S_\pi(x, y^*) \geq 0$ indicates whether the model output matches the correct answer. In the multiple-choice questions setting, S_π is the accuracy and thus is 1 if the predicted option equals y^*, and 0 otherwise. In the open-ended generation setting, S_π is a score determined by an external LLM judge.



(a) Egypt localized

(b) Nepal non-localized

Figure 2: Localized (a) and non-localized (b) training and evaluation loss comparison between BiPO and YaPO for Egypt (a) and Nepal (b).

Let $\bar{p} = \frac{1}{2}(p_{\text{loc}} + p_{\text{non}})$. *The* performance–normalized localization gap *is:*

$$\text{PNLG}_\alpha(\pi) = \mathbb{E}_{(x_{\text{loc}}, x_{\text{nonloc}}, y^*)\sim\mathcal{D}} \left[\frac{p_{\text{loc}} - p_{\text{non}}}{\bar{p}^{\,\alpha} + \varepsilon} \right], \tag{5}$$

with $\varepsilon > 0$ *arbitrarily small for numerical stability and* $\alpha \in [0, 1]$ *controlling the strength of the normalization.*

Definition 2 (Robust Cultural Accuracy (RCA))**.** *Using the same notation, the* robust cultural accuracy *is the harmonic mean of localized and non–localized accuracies:*

$$\text{RCA}(\pi) = \mathbb{E}_{(x_{\text{loc}}, x_{\text{nonloc}}, y^*)\sim\mathcal{D}} \left[\frac{2\, p_{\text{loc}}\, p_{\text{non}}}{p_{\text{loc}} + p_{\text{non}} + \varepsilon} \right]. \tag{6}$$

with $\varepsilon > 0$ *arbitrarily small for numerical stability.*

Design choice of metrics. A raw localization gap $p_{\text{loc}} - p_{\text{non}}$ can be misleading: a weak model may display a small gap simply because both accuracies are near zero. PNLG corrects for this by normalizing the gap with the mean performance $\bar{p}$, so models with trivially low accuracy are penalized. RCA complements this by rewarding methods that are both accurate and balanced across localized and non–localized prompts. Together, PNLG and RCA provide a more faithful evaluation of cultural alignment than raw gap alone.

Remark 1. *Different values of* α *control the strength of the normalization:*

- $\alpha = 0$*: reduces to the raw gap* $p_{\text{loc}} - p_{\text{non}}$*, without any performance normalization.*
- $\alpha = 0.5$*: applies a moderate normalization, balancing sensitivity to both the gap and the average performance.*
- $\alpha = 1$*: normalizes by the mean performance* $\bar{p}$*, strongly penalizing models that show small gaps only because both localized and non-localized accuracies are very low (**our default choice**).*

Baselines. We benchmark the performances of YaPO against two baselines:

1. **No steering**: the original Gemma-2-2B model without any intervention.
2. **BiPO** (Cao et al., 2024): which optimizes dense steering vectors directly in the residual stream via bi-directional preference optimization.

These baselines allow us to disentangle the contributions of sparse representations and preference optimization in improving cultural alignment.

3.2 Training Dynamics Analysis

We begin by comparing the training dynamics of YaPO and BiPO. Empirically, we find that the same behavior occur for all countries and scenarios. Thus, for conciseness matters, we report training and

evaluation loss logs for "Egypt" and "Nepal" under both the "localized" and "non-localized" cultural adaptation settings. Figures 2a–2b show training and evaluation loss over optimization steps for both methods (YaPO and BiPO).

The contrast is striking: **YaPO converges an order of magnitude faster**, with loss consistently dropping below 0.1 in under than 150 steps in both scenarios, whereas BiPO remains above 0.3 even after 600 steps. This rapid convergence stems from and underscores the advantage of operating in the sparse SAE latent space, where disentangled features yield cleaner gradients and more stable optimization. Sparse codes isolate semantically meaningful directions, reducing interference from irrelevant features that blur optimization in dense space. In contrast, BiPO remains tied to the dense residual space, where multi-semanticity and superposition entangle behavioral factors, hindering convergence, stability, and interpretability, particularly in tasks that require disentangling closely related features.

4 EVALUATION

We evaluate YaPO againt BiPO and the baseline model without steering on our curated multilingual cultural adaptation benchmark using both multiple-choice questions (MCQs) and open-ended generation (OG). To assess absolute alignment as well as robustness to the explicit–implicit localization gap, we consider the three settings: localized, non-localized, and mixed prompts. MCQ performance is measured by accuracy[1], while OG responses are scored by an external LLM judge for consistency with the gold answer (see Appendix **??** for the evaluation prompt).

4.1 MULTIPLE-CHOICE QUESTIONS

Table 1 summarizes MCQ results by language and country.

Quantitative analysis. Across all languages and settings, YaPO consistently surpasses BiPO and the baseline model. Country-level peaks include Nepal at 70.4% (YaPO), 44.9% (BiPO) in localized prompts (+26.7 over baseline) and 68.2% (YaPO), 40.7% (BiPO) in non-localized (+30.7% over baseline), demonstrating that sparse steering vectors excel even under implicit cultural cues. In particular, we find that non-localized prompts benefit the most, confirming that sparse preference-optimized vectors generalize better under implicit cues, whereas BiPO often collapses by entangling neighboring cultural features. Additionally, we observe that the performance improvement is stable and consistent throughout the epochs for YaPO while BiPO see Figure 5. In general, we also observe that the strongest improvements occur in the highest-resource language of each group, where pretraining exposure is richer and steering at inference time suffices to align outputs.

Qualitative analysis. Training dynamics analysis in Section 3.2 showed that YaPO stabilizes rapidly. We observe this at inference time on the test data. In the MCQ setting, once the correct answer is identified, it remains consistent across epochs with only minor token-level variations, whereas BiPO frequently overwrites correct predictions and occasionally shifts to an incorrect language. In the open-ended setting, BiPO exhibits substantial drift in generated outputs across epochs, while YaPO remains comparatively stable. This stability underlines YaPO's advantage in benefiting from robust and interpretable steering.

4.2 OPEN-ENDED GENERATION

Table 2 summarizes OG results by language and country.

Quantitative analysis. We observe consistent improvements from YaPO over both baseline and BiPO in most settings. In localized prompts, YaPO yields steady gains (e.g., +0.40 in English and +0.81 in Hindi on average), with the largest improvements again concentrated in lower-resource languages such as Hindi and Arabic. Under non-localized prompts, YaPO provides further advantages, especially in Spanish (+0.53) and Hindi (+0.19), confirming its robustness when cultural cues are implicit. While BiPO occasionally achieves higher scores in isolated cases (e.g.,

[1]The ground-truth answer is annotated using a `\boxed{k}` tag, where k denotes the index of the correct choice.

Table 1: Multiple-Choice Questions Performance by Language and Country across settings.

		Localized			Non-localized			Both		
Language	**Country**	**Baseline**	**BiPO**	**YaPO (ours)**	**Baseline**	**BiPO**	**YaPO (ours)**	**Baseline**	**BiPO**	**YaPO (ours)**
English	UK	36.4%	36.8% (+0.4%)	**49.1%** (+12.7%)	23.2%	30.3% (+7.1%)	**39.1%** (+15.9%)	29.0%	33.8% (+4.8%)	**43.6%** (+14.6%)
	USA	45.5%	51.9% (+6.4%)	**59.8%** (+14.3%)	40.2%	45.9% (+5.7%)	**54.4%** (+14.2%)	44.7%	45.2% (+0.5%)	**57.5%** (+12.8%)
	Australia	48.2%	51.1% (+2.9%)	**59.8%** (+11.6%)	23.8%	31.1% (+7.3%)	**38.8%** (+15.0%)	33.3%	37.9% (+4.6%)	**50.2%** (+16.9%)
	Average	*43.4%*	*46.6%* (+3.2%)	***56.2%*** (+12.9%)	*29.1%*	*35.8%* (+6.7%)	***44.1%*** (+15.0%)	*35.7%*	*39.0%* (+3.3%)	***50.4%*** (+14.7%)
Spanish	Bolivia	22.8%	29.4% (+6.6%)	**42.1%** (+19.3%)	14.5%	17.4% (+2.9%)	**24.6%** (+10.1%)	18.5%	25.3% (+6.8%)	**35.5%** (+17.0%)
	Mexico	24.4%	22.5% (–1.9%)	**35.2%** (+10.8%)	13.3%	18.4% (+5.1%)	**27.2%** (+13.9%)	18.6%	21.2% (+2.6%)	**30.0%** (+11.4%)
	Spain	46.5%	50.8% (+4.3%)	**61.6%** (+15.1%)	31.8%	35.1% (+3.3%)	**43.5%** (+11.7%)	**37.3%**	41.1% (+3.8%)	**52.3%** (+15.0%)
	Average	*31.2%*	*34.2%* (+3.0%)	***46.3%*** (+15.1%)	*19.9%*	*23.6%* (+3.7%)	***31.8%*** (+11.9%)	***24.8%***	*29.2%* (+4.4%)	***39.3%*** (+14.5%)
Portuguese	Brazil	23.4%	27.9% (+4.5%)	**41.6%** (+18.2%)	17.7%	22.2% (+4.5%)	**34.8%** (+17.1%)	19.9%	27.3% (+7.4%)	**39.1%** (+19.2%)
	Mozambique	21.8%	28.0% (+6.2%)	**37.2%** (+15.4%)	19.3%	25.7% (+6.4%)	**27.5%** (+8.2%)	20.2%	25.0% (+4.8%)	**32.1%** (+11.9%)
	Portugal	33.5%	37.6% (+4.1%)	**53.2%** (+19.7%)	28.7%	35.2% (+6.5%)	**52.3%** (+23.6%)	**32.2%**	34.5% (+2.3%)	**54.0%** (+21.8%)
	Average	*26.2%*	*31.2%* (+5.0%)	***44.0%*** (+17.8%)	*21.9%*	*27.7%* (+5.8%)	***38.2%*** (+16.3%)	***24.1%***	*28.9%* (+4.8%)	***41.7%*** (+17.6%)
Arabic	Egypt	43.1%	45.1% (+2.0%)	**47.7%** (+4.6%)	36.0%	39.8% (+3.8%)	**43.6%** (+7.6%)	36.1%	42.2% (+6.1%)	**50.2%** (+14.1%)
	KSA	16.1%	19.9% (+3.8%)	**20.2%** (+4.1%)	16.7%	18.9% (+2.2%)	**19.2%** (+2.5%)	17.1%	19.5% (+2.4%)	**20.9%** (+3.8%)
	Levantine	15.0%	16.9% (+1.9%)	**16.9%** (+1.9%)	10.3%	11.4% (+1.1%)	**13.1%** (+2.8%)	12.4%	14.6% (+2.2%)	**15.3%** (+2.9%)
	Morocco	12.6%	13.6% (+1.0%)	**14.0%** (+1.4%)	12.6%	13.6% (+1.0%)	**14.0%** (+1.4%)	11.6%	**13.8%** (+2.2%)	13.6% (+2.0%)
	Average	*21.7%*	*23.9%* (+2.2%)	***24.7%*** (+3.0%)	*21.0%*	*23.4%* (+2.4%)	***22.5%*** (+3.5%)	*19.3%*	*22.5%* (+3.2%)	***25.0%*** (+5.7%)
Hindi	India	21.6%	23.4% (+1.8%)	**41.1%** (+19.5%)	22.2%	26.1% (+3.9%)	**39.9%** (+17.7%)	20.3%	22.4% (+2.1%)	**42.9%** (+22.6%)
	Nepal	43.7%	44.9% (+1.2%)	**70.4%** (+26.7%)	37.0%	40.7% (+3.7%)	**68.2%** (+31.2%)	41.6%	42.1% (+0.5%)	**70.6%** (+29.0%)
	Average	*32.7%*	*34.2%* (+1.5%)	***55.8%*** (+23.1%)	*29.6%*	*33.4%* (+3.8%)	***54.1%*** (+24.5%)	*31.0%*	*32.3%* (+1.3%)	***56.8%*** (+25.8%)

Portuguese–Mozambique), these improvements are unstable and sometimes degrade overall performance (–3.83 localized, –0.35 both). On average, YaPO achieves the best trade-off across all evaluation modes, reaching 6.01 in localized, 0.61 in non-localized, and 1.16 in the combined setting.

Qualitative analysis. Similar to the behavior observed in MCQs.

Table 2: Open-Ended Performance by Language and Country across settings.

		Localized			Non-localized			Both		
Language	**Country**	**Baseline**	**BiPO**	**YaPO (ours)**	**Baseline**	**BiPO**	**YaPO (ours)**	**Baseline**	**BiPO**	**YaPO (ours)**
English	UK	6.39	6.54 (+0.15)	**6.63** (+0.24)	0.83	1.41 (+0.58)	**1.48** (+0.65)	0.83	**0.98** (+0.15)	0.95 (+0.12)
	USA	6.78	7.14 (+0.36)	**7.20** (+0.42)	0.59	1.30 (+0.71)	**1.90** (+1.31)	0.66	1.56 (+0.90)	**1.89** (+1.23)
	Australia	6.78	7.10 (+0.32)	**7.32** (+0.54)	0.68	1.44 (+0.76)	**1.95** (+1.27)	1.93	**2.71** (+0.78)	2.31 (+0.38)
	Average	*6.65*	*6.93* (+0.28)	***7.05*** (+0.40)	*0.70*	*1.38* (+0.68)	***1.78*** (+1.08)	*1.14*	***1.75*** (+0.61)	*1.72* (+0.58)
Spanish	Spain	6.68	**6.79** (+0.11)	6.77 (+0.09)	2.76	2.80 (+0.04)	**2.99** (+0.23)	2.79	**3.28** (+0.49)	2.96 (+0.17)
	Mexico	6.67	6.69 (+0.02)	**6.94** (+0.27)	2.38	2.98 (+0.60)	**3.04** (+0.66)	2.38	**3.04** (+0.66)	2.98 (+0.60)
	Bolivia	6.57	6.65 (+0.08)	**6.74** (+0.17)	1.86	2.14 (+0.28)	**2.26** (+0.40)	1.86	**2.26** (+0.40)	2.14 (+0.28)
	Average	*6.64*	*6.71* (+0.07)	***6.82*** (+0.18)	*2.33*	*2.64* (+0.31)	***2.76*** (+0.43)	*2.34*	***2.86*** (+0.52)	*2.69* (+0.35)
Portuguese	Brazil	6.61	6.67 (+0.06)	**6.70** (+0.09)	0.76	0.80 (+0.04)	**0.79** (+0.03)	2.70	2.65 (–0.05)	**2.90** (+0.20)
	Mozambique	6.36	**6.47** (+0.11)	6.44 (+0.08)	0.36	**0.65** (+0.29)	0.52 (+0.16)	2.98	2.63 (–0.35)	**2.78** (+0.20)
	Portugal	6.49	6.60 (+0.11)	**6.66** (+0.17)	1.30	**1.70** (+0.40)	1.53 (+0.23)	0.47	**0.60** (+0.13)	**0.60** (+0.13)
	Average	*6.49*	*6.58* (+0.09)	***6.60*** (+0.11)	*0.81*	*1.05* (+0.24)	***1.11*** (+0.30)	*2.05*	*2.29* (+0.24)	***2.43*** (+0.38)
Arabic	Egypt	5.08	**5.47** (+0.39)	5.11 (+0.03)	0.73	1.00 (+0.27)	**1.08** (+0.35)	1.76	**2.08** (+0.32)	1.97 (+0.21)
	KSA	3.95	4.22 (+0.27)	**5.34** (+1.39)	0.96	1.10 (+0.14)	**1.37** (+0.41)	0.44	0.46 (+0.02)	**0.61** (+0.17)
	Levantine	3.85	3.90 (+0.05)	**4.80** (+0.95)	0.21	0.42 (+0.21)	**0.56** (+0.35)	0.78	0.79 (+0.01)	**1.16** (+0.38)
	Morocco	3.48	3.46 (–0.02)	**4.18** (+0.70)	0.65	0.77 (+0.12)	**0.86** (+0.21)	0.58	0.61 (+0.03)	**0.74** (+0.16)
	Average	*4.09*	*4.26* (+0.17)	***4.86*** (+0.77)	*0.64*	*0.82* (+0.19)	***0.97*** (+0.33)	*0.89*	*0.99* (+0.10)	***1.12*** (+0.23)
Hindi	India	5.29	5.55 (+0.26)	**6.11** (+0.82)	0.11	0.39 (+0.28)	**0.43** (+0.32)	0.58	**1.03** (+0.45)	0.93 (+0.35)
	Nepal	5.11	5.25 (+0.14)	**5.90** (+0.79)	0.73	0.75 (+0.02)	**0.79** (+0.06)	1.07	1.25 (+0.18)	**1.38** (+0.31)
	Average	*5.20*	*5.40* (+0.20)	***6.01*** (+0.81)	*0.42*	*0.57* (+0.15)	***0.61*** (+0.19)	*0.83*	*1.14* (+0.32)	***1.16*** (+0.33)

4.3 Explicit–Implicit Localization Gap

Table 3 reports RCA and PNLG. We recall that RCA (eq.6) is the harmonic mean of localized and non-localized accuracies, thus rewarding models that are both accurate and balanced across settings. High RCA therefore indicates robust cultural competence rather than overfitting to explicit prompts. In contrast, PNLG (eq.5) measures the relative difference between localized and non-localized performance, normalized by their average; lower PNLG implies a smaller explicit–implicit localization gap.

Overall, YaPO achieves the best trade-off: it substantially increases RCA (41.2%, +14.5% over baseline) while also reducing PNLG (0.184, –27.3% over baseline). This shows that YaPO not only improves absolute cultural accuracy but also yields more consistent behavior across prompt types, whereas BiPO tends to reduce the gap without comparable gains in robustness.

Table 3: RCA and PNLG Analysis by Language for MCQ and Open-Ended Tasks

Language	RCA ↑ (Higher is better)						PNLG ↓ (Lower is better)					
	MCQ (%)			Open-Ended (0-10 scale)			MCQ			Open-Ended		
	Base	BiPO	YaPO	Base	BiPO	YaPO	Base	BiPO	YaPO	Base	BiPO	YaPO
Arabic	20.1	22.2 (↑10.4%)	**23.5** (↑16.9%)	1.08	1.36 (↑25.9%)	**1.60** (↑48.1%)	0.129	0.141 (↑9.3%)	**0.098** (↓24.0%)	1.470	1.359 (↓7.6%)	**1.346** (↓8.4%)
English	34.3	40.2 (↑17.2%)	**49.2** (↑43.4%)	1.26	2.30 (↑82.5%)	**2.84** (↑125.4%)	0.415	0.268 (↓35.4%)	**0.249** (↓40.0%)	1.618	1.333 (↓17.6%)	**1.198** (↓26.0%)
Hindi	31.0	33.7 (↑8.7%)	**54.9** (↑77.1%)	0.75	1.02 (↑36.0%)	**1.10** (↑46.7%)	0.069	**-0.005** (↓107.2%)	0.031 (↓55.1%)	1.709	1.619 (↓5.3%)	**1.632** (↓4.5%)
Portuguese	23.8	29.3 (↑23.1%)	**40.8** (↑71.4%)	1.40	1.77 (↑26.4%)	**1.62** (↑15.7%)	0.184	**0.126** (↓31.5%)	0.165 (↓10.3%)	1.569	1.462 (↓6.8%)	**1.511** (↓3.7%)
Spanish	24.2	27.9 (↑15.3%)	**37.6** (↑55.4%)	3.44	3.78 (↑9.9%)	**3.92** (↑14.0%)	0.470	**0.360** (↓23.4%)	0.375 (↓20.2%)	0.965	0.875 (↓9.3%)	**0.851** (↓11.8%)
Overall	26.7	30.7 (↑15.0%)	**41.2** (↑54.3%)	1.59	2.05 (↑28.9%)	**2.22** (↑39.6%)	0.253	**0.178** (↓29.6%)	0.184 (↓27.3%)	1.466	1.330 (↓9.3%)	**1.308** (↓10.8%)

4.4 Generalization to Other Domains

To assess whether cultural steering vectors specialize too narrowly, we evaluate them on BiPO's benchmarks. On the hallucination task, the baseline model reaches a score of 1.580, while BiPO slightly underperforms at 1.575. In contrast, YaPO achieves a score of **1.680**, clearly outperforming both. This demonstrates that learning in sparse space not only improves cultural alignment but also **generalizes** to broader alignment dimensions such as hallucination reduction.

5 Related Works

Alignment and controllability. RLHF (Christiano et al., 2017; Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022) has become the standard approach to align LLMs, training a reward model on human preference data and fine-tuning with PPO (Schulman et al., 2017) under the Bradley-Terry framework (Bradley & Terry, 1952). Recent methods simplify this pipeline by bypassing explicit reward modeling: DPO (Rafailov et al., 2024) directly optimizes on preference pairs, while SLiC (Zhao et al., 2023) introduces a contrastive calibration loss with regularization toward the SFT model. Statistical Rejection Sampling (Liu et al., 2024) unifies both objectives and provides a tighter policy estimate.

Activation Engineering. steers LLMs by freezing weights and perturbing activations in hidden layers. Early work showed that sentence-specific vectors could be optimized to reproduce target text (Subramani et al., 2022), though this required costly per-sample gradient descent. Activation addition (Turner et al., 2023) instead computes activation differences between prompt pairs, but its performance is inconsistent. CAA (Rimsky et al., 2023) averages across many preference pairs, and has been applied to steer personas and mitigate hallucinations in LLaMA-2, while Wang & Shu (2023) extended this to free-form prompts, even demonstrating safety compromises. However, because these vectors are derived directly from prompt activations, they often fail to reflect the model's actual generation behavior, particularly in alignment-critical cases. Beyond MLP activations, other approaches perturb attention heads: Li et al. (2024) shifted truth-correlated heads to improve factuality, and Liu et al. (2023) replaced demonstrations in in-context learning with latent activation shifts. Overall, existing activation-based methods remain noisy and unstable. Recently, BiPO (Cao et al., 2024) reframed steering as preference optimization, directly learning dense vectors with a bi-directional DPO loss, and directly optimizes steering vectors in the activation space rather than model weights, yielding more accurate, disentangled, and controllable representations of target behaviors.

Sparse activation steering. To address superposition, Sparse Autoencoders (SAEs) (Lieberum et al., 2024) decompose activations into high-dimensional sparse codes that approximate monosemantic features. Sparse Activation Steering (SAS) (Bayat et al., 2025) operationalized this for behavior control, building steering vectors by averaging sparse activations from contrastive datasets. SAS achieves interpretable, compositional, and fine-grained control, while preserving general utility under moderate steering. However, because its sparse directions are not optimized against preferences, its effectiveness remains limited compared to preference-optimized methods.

Positioning of YaPO. BiPO provides strong optimization but suffers from dense entanglement; SAS offers interpretability but lacks optimization. YaPO unifies these lines by learning sparse, preference-optimized steering vectors in SAE space. This yields disentangled, interpretable, and stable steering, with improved convergence and generalization across cultural alignment, truthfulness, hallucination suppression, and jailbreak defense.

6 Conclusion

In this work, we introduced **YaPO**, a reference-free method that learns sparse, preference-optimized steering vectors in the latent space of Sparse Autoencoders. Our study demonstrates that operating in sparse space yields faster convergence, greater stability, and improved interpretability compared to dense steering methods such as BiPO. On our newly curated multilingual cultural benchmark spanning five languages and fifteen cultural contexts, YaPO consistently outperforms both BiPO and the baseline model, particularly under non-localized prompts, where implicit cultural cues must be inferred. Beyond culture, YaPO generalizes to other alignment dimensions such as hallucination mitigation, underscoring its potential as a general recipe for efficient and fine-grained alignment.

7 Limitations

While promising, YaPO also comes with certain limitations. First, our evaluation is restricted to **Gemma-2-2B**; although we also observe the same trend with **Gemma-2-9B** in training (details omitted for brevity) further work is required to test whether the observed gains scale to larger models or transfer across architectures. Second, our generalization study was limited in scope: although we reported improvements on hallucination, other alignment dimensions such as truthfulness and jailbreak resistance were not explored in depth. Finally, while sparse vectors improve interpretability, a systematic analysis of what individual features encode was beyond the scope of this work, though this can be done via encoding BiPO's dense steering vector and computing the difference with YaPO's steering vector, and looking at the features misalignment.

8 LLM Usage

We use LLMs solely to polish writing and clarify ideas, keeping all scientific reasoning human-driven. The model acts only as a stylistic assistant, enhancing readability without contributing content.

References

Reza Bayat, Ali Rahimi-Kalahroudi, Mohammad Pezeshki, Sarath Chandar, and Pascal Vincent. Steering large language model activations in sparse spaces, 2025. URL https://arxiv.org/abs/2503.00177.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Yuanpu Cao, Tianrong Zhang, Bochuan Cao, Ziyi Yin, Lu Lin, Fenglong Ma, and Jinghui Chen. Personalized steering of large language models: Versatile steering vectors through bi-directional preference optimization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=7qJFkuZdYo.

Paul F Christiano, Jan Leike, Tom B Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. *arXiv preprint arXiv:2411.08745*, 2024.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition, 2022. URL https://arxiv.org/abs/2209.10652.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. *arXiv preprint arXiv:2401.06102*, 2024.

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

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2, 2024. URL https://arxiv.org/abs/2408.05147.

Shuchen Liu, Long Xing, and James Zou. In-context vectors: Making in context learning more effective and controllable through latent space steering. *arXiv preprint arXiv:2311.06668*, 2023.

Tianyu Liu, Yizhe Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jing Liu. Statistical rejection sampling improves preference optimization. In *International Conference on Learning Representations*, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition, 2024. URL https://arxiv.org/abs/2312.06681.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Nathan Rimsky, Noam Gabrieli, Jonas Schulz, Matthew Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Carson Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3008–3021, 2020.

Nishanth Subramani, Nishanth Suresh, and Matthew E Peters. Extracting latent steering vectors from pretrained language models. *arXiv preprint arXiv:2205.05124*, 2022.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024. URL `https://arxiv.org/abs/2408.00118`.

Alexander Matt Turner, Leon Thiergart, David Udell, Geoffrey Leech, Ulisse Mini, and Michael MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

Veniamin Veselovsky, Berke Argin, Benedikt Stroebl, Chris Wendler, Robert West, James Evans, Thomas L. Griffiths, and Arvind Narayanan. Localized cultural knowledge is conserved and controllable in large language models, 2025. URL `https://arxiv.org/abs/2504.10191`.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401, 2020.

Haoran Wang and Kai Shu. Backdoor activation attack: Attack large language models using activation steering for safety-alignment. *arXiv preprint arXiv:2311.09433*, 2023.

Yizhe Zhao, Rishabh Joshi, Tianyu Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*, 2023.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A Training Details

We summarize the training configuration and hyperparameters in Table 4.

Table 4: Training configuration and hyperparameters.

Hardware	Single node with 8 × AMD MI210 GPUs
Epochs	20
Batch size	4 (gradient accumulation = 1)
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.999$) with weight decay of 0.05
Learning rate	5×10^{-4}
LR scheduler	Cosine decay with 100 warmup steps
Max prompt length	512 tokens
Max new tokens	2048
SAE layer	15
SAE vector size	65k
Average index (layer 15)	68

B Layer Discovery

We employ *activation patching* (Ghandeharioun et al., 2024; Dumas et al., 2024; Vig et al., 2020) to identify which layers of the LLM contribute most strongly to cultural localization. In our setting, the *slocalized prompt* $x_{\text{localized}}$ is the localized version of the input (e.g., specifying the country or culture), whereas the *non-localized prompt* x_{nonloc} is the -localized variant (e.g., without cultural specification).

Due to causal masking in the attention layers, the latent representation of the i-th input token after the j-th transformer block depends on all preceding tokens:

$$h_i^{(j)} = h_i^{(j)}(x_1, \ldots, x_i).$$

For clarity, we omit this explicit dependence when clear from context and use the shorthand notation $h^{(j)}(x)_i$.

We first perform a forward pass on the localized (source) prompt and extract its latent representation $h_i^{(j)}(x_{\text{localized}})$ at each layer. During the forward pass on the non-localized (target) prompt, we *patch* its latent representation by overwriting $h_i^{(j)}(x_{\text{nonloc}})$ with the localized one, producing a perturbed forward pass $\tilde{P}(x_{\text{nonloc}})$. By comparing $\tilde{P}(x_{\text{nonloc}})$ to the original prediction $P(x_{\text{nonloc}})$, we quantify how much information from each layer of the localized prompt contributes to aligning the model's behavior with the culturally appropriate response.

Concretely, for our analysis we focus on the latent representation at the last token position $t_{\text{localized}}$ in the localized prompt, i.e.,

$$h_{t_{\text{localized}}}^{(j)}(x_{\text{localized}}),$$

and patch this into the corresponding position in the target forward pass. Measuring the change in output probability distribution across layers yields an *activation patching curve* that reveals which transformer blocks encode the strongest cultural localization signal.and we do this for two countries for specific language so we choose two countries Egypt and Morocco and the data was just question that are loclaized and non localized and then we have for the answers the for egypt the egyptian answer and for Morocco we have moroccon answer and western answer and then we apply the activation batching on both as we defined in the above so that we can find the layers, for Gemma models for both Gemma-2 9b, and Gemma-2 2b, and as we see in the figure 2

C More Evaluations

Figure 5 summarizes the aggregated evaluation results. Across all languages and cultural contexts, YaPO consistently achieves higher accuracy and lower variance compared to BiPO. This demonstrates not only improved performance but also greater training stability, highlighting the advantage of learning sparse steering vectors in the SAE latent space.

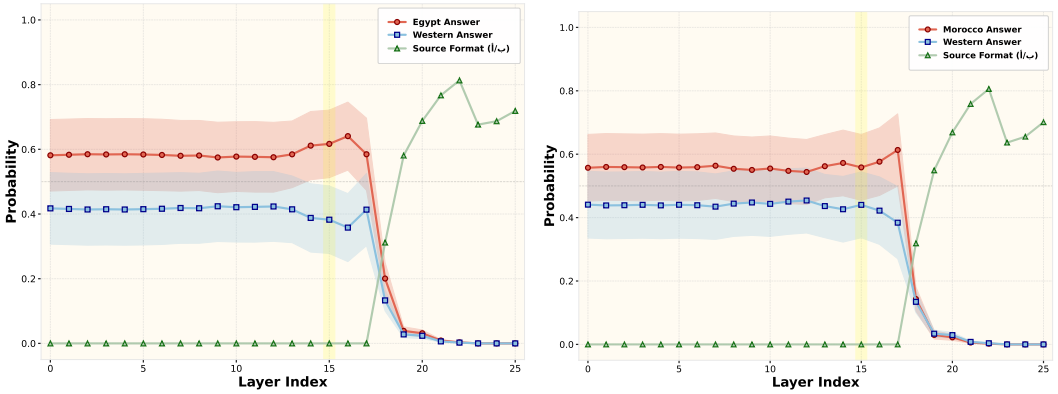


Figure 3: Activation patching analysis on **Gemma-2-2B**. We intervene across layers to trace cultural features in model representations. The plots show the probability of producing culturally specific answers (Egypt, Morocco) versus Western defaults as activations are patched. We empirically identify **layer 15** as the most culturally relevant layer.

Table 5: PNLG and RCA Analysis by Country for MCQ and Open-Ended Tasks

Language	Country	Task	PNLG ↓ (Lower is better)			RCA ↑ (Higher is better)		
			Baseline	BiPO	YaPO (ours)	Baseline	BiPO	YaPO (ours)
English	UK	MCQ	0.443	**0.194** (↓56.2%)	0.227 (↓48.8%)	28.3%	33.2% (↑4.9%)	**43.5%** (↑15.2%)
		Open	1.540	1.291 (↓16.2%)	**1.270** (↓17.5%)	14.7%	23.2% (↑8.5%)	**24.2%** (↑9.5%)
	USA	MCQ	0.124	0.123 (↓0.8%)	**0.095** (↓23.4%)	42.7%	48.7% (↑6.0%)	**57.0%** (↑14.3%)
		Open	1.680	1.384 (↓17.6%)	**1.165** (↓30.7%)	10.9%	22.0% (↑11.1%)	**30.1%** (↑19.2%)
	Australia	MCQ	0.678	0.487 (↓28.2%)	**0.426** (↓37.2%)	31.9%	38.7% (↑6.8%)	**47.1%** (↑15.2%)
		Open	1.635	1.326 (↓18.9%)	**1.159** (↓29.1%)	12.4%	23.9% (↑11.5%)	**30.8%** (↑18.4%)
Spanish	Spain	MCQ	0.375	0.366 (↓2.4%)	**0.344** (↓8.3%)	37.8%	41.5% (↑3.7%)	**51.0%** (↑13.2%)
		Open	0.831	0.832 (↑0.1%)	**0.775** (↓6.7%)	39.1%	39.6% (↑0.5%)	**41.5%** (↑2.4%)
	Mexico	MCQ	0.589	**0.200** (↓66.0%)	0.256 (↓56.5%)	17.2%	20.2% (↑3.0%)	**30.7%** (↑13.5%)
		Open	0.948	**0.767** (↓19.1%)	0.782 (↓17.5%)	35.1%	41.2% (↑6.1%)	**42.3%** (↑7.2%)
	Bolivia	MCQ	**0.445**	0.513 (↑15.3%)	0.525 (↑18.0%)	17.7%	21.9% (↑4.2%)	**31.1%** (↑13.4%)
		Open	1.117	1.026 (↓8.1%)	**0.996** (↓10.8%)	29.0%	32.4% (↑3.4%)	**33.8%** (↑4.8%)
Portuguese	Brazil	MCQ	0.277	0.228 (↓17.7%)	**0.178** (↓35.7%)	20.2%	24.7% (↑4.5%)	**37.9%** (↑17.7%)
		Open	1.588	1.572 (↓1.0%)	**1.578** (↓0.6%)	13.6%	14.3% (↑0.7%)	**14.1%** (↑0.5%)
	Mozambique	MCQ	0.122	**0.086** (↓29.5%)	0.300 (↑145.9%)	20.5%	26.8% (↑6.3%)	**31.6%** (↑11.1%)
		Open	1.786	1.635 (↓8.5%)	**1.701** (↓4.8%)	6.8%	11.8% (↑5.0%)	**9.6%** (↑2.8%)
	Portugal	MCQ	0.154	0.066 (↓57.1%)	**0.017** (↓89.0%)	30.9%	36.4% (↑5.5%)	**52.7%** (↑21.8%)
		Open	1.332	1.181 (↓11.3%)	**1.253** (↓5.9%)	21.7%	27.0% (↑5.3%)	**24.9%** (↑3.2%)
Arabic	Egypt	MCQ	0.180	0.125 (↓30.6%)	**0.090** (↓50.0%)	39.2%	42.3% (↑3.1%)	**45.6%** (↑6.4%)
		Open	1.497	1.382 (↓7.7%)	**1.302** (↓13.0%)	12.8%	16.9% (↑4.1%)	**17.8%** (↑5.0%)
	KSA	MCQ	**-0.037**	0.052 (↑240.5%)	0.051 (↑237.8%)	16.4%	19.4% (↑3.0%)	**19.7%** (↑3.3%)
		Open	**1.218**	1.173 (↓3.7%)	1.183 (↓2.9%)	15.4%	17.5% (↑2.1%)	**21.8%** (↑6.4%)
	Levantine	MCQ	0.372	0.389 (↑4.6%)	**0.253** (↓32.0%)	12.2%	13.6% (↑1.4%)	**14.8%** (↑2.6%)
		Open	1.793	1.611 (↓10.2%)	**1.582** (↓11.8%)	4.0%	7.6% (↑3.6%)	**10.0%** (↑6.0%)
	Morocco	MCQ	**0.000**	0.000 (0.0%)	0.000 (0.0%)	12.6%	13.6% (↑1.0%)	**14.0%** (↑1.4%)
		Open	1.370	**1.272** (↓7.2%)	1.317 (↓3.9%)	11.0%	12.6% (↑1.6%)	**14.3%** (↑3.3%)
Hindi	India	MCQ	**-0.027**	-0.109 (↓303.7%)	0.030 (↑211.1%)	21.9%	24.7% (↑2.8%)	**40.5%** (↑18.6%)
		Open	1.919	**1.737** (↓9.5%)	**1.737** (↓9.5%)	2.2%	7.3% (↑5.1%)	**8.0%** (↑5.8%)
	Nepal	MCQ	0.166	0.098 (↓41.0%)	**0.032** (↓80.7%)	40.1%	42.7% (↑2.6%)	**69.3%** (↑29.2%)
		Open	**1.500**	**1.500** (0.0%)	1.528 (↑1.9%)	12.8%	13.1% (↑0.3%)	**13.9%** (↑1.1%)

D DATASET

Our dataset is explicitly designed to make these failures measurable by stress-testing *implicit vs. explicit* cultural localization under *within-language* control. We cover 52 lived-experience topics (Table 6) meals, routines, family relations, greetings and etiquette, financial habits, ceremonies and

Evaluate this model response (Score 1–10) based on Language Quality, Factual Accuracy, Cultural Appropriateness, and Content Quality.

1) Language Quality (0–2):
- 0: Nonsensical text (made-up words, broken grammar, gibberish)
- 1: Poor language with many errors
- 2: Correct, natural language

2) Factual Accuracy (0–3):
- 0: Fabricated information (fake practices/customs that don't exist)
- 1: Mix of correct and incorrect facts
- 2: Mostly factually correct
- 3: Completely accurate information

3) Cultural Appropriateness (0–3):
- 0: Wrong culture or generic/Western response
- 1: Generic response, not country-specific
- 2: Mostly appropriate for target culture
- 3: Perfect cultural accuracy for specific country

4) Content Quality (0–2):
- 0: Severe issues (excessive repetition, refusal, off-topic)
- 1: Partial answer with problems
- 2: Complete, relevant, well-structured answer

Question: {question[:]}
Expected Answer: {ground_truth[:]}
Model Response ({response_type}): {model_response[:]}

CRITICAL CHECKS
Are cultural practices authentic (not fabricated)?
Is content non-repetitive?
Is it culturally specific (not generic)?
Does it align with the expected answer?

Figure 4: Evaluation prompt

mourning, holidays, childhood games, music and idioms, because these domains reveal *norms* rather than trivia. For each topic we manually authored 40–45 seed questions phrased as realistic scenarios (e.g., weekend breakfast, commute habits, hospitality customs). Every question appears in *paired form*: a *localized* variant that names the country and a *non-localized* variant that omits it, forcing the model to rely on dialect and situational cues. Each item is cast as a multiple-choice question with *one culturally valid option per country* within the same language group, written in that country's *dialect*, plus a *Western control option* expressed in a standardized register (MSA for Arabic) to isolate culture from translation artifacts. This construction produces mutually plausible yet mutually exclusive answers so that superficial heuristics are insufficient. It enables principled measurement of the *Localization Gap* (accuracy shift from non-localized to localized form), *Intra-language Dominance Bias* (systematic preference for one country in non-localized form), and *Stereotype Preference*

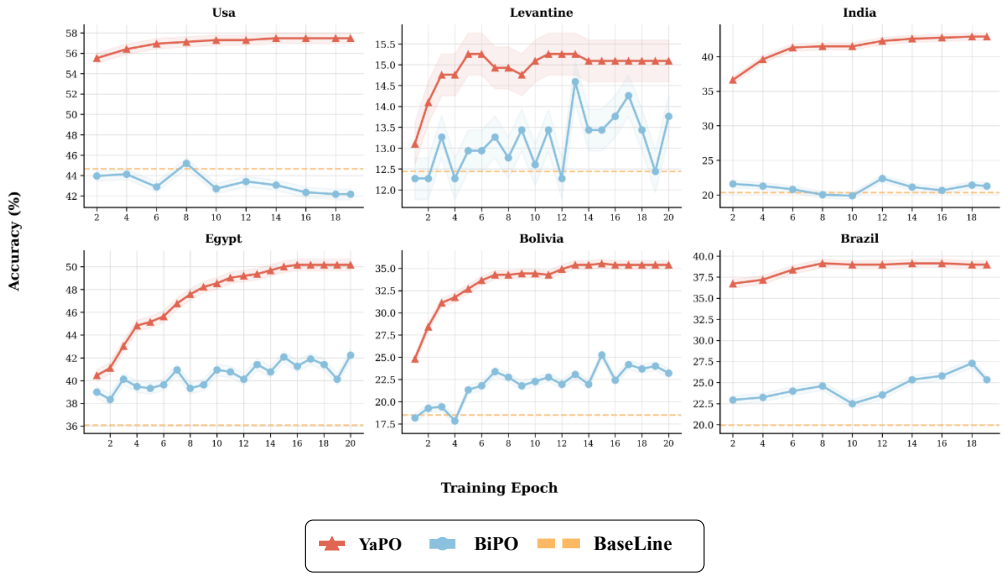


Figure 5: YaPO consistently improves performance and exhibits greater stability compared to BiPO.

(gravitating toward caricatured or Western answers against human-majority ground truth). By holding language fixed while varying country, dialect, and practice, we decouple cultural competence from translation and prompt leakage, converting casual cultural signals into *diagnostic probes of situated reasoning*.

D.1 Data Curation Pipeline

We built the dataset through a multi-stage pipeline (see Figure **??**) that integrates generation, filtering, and contrastive packaging. We began by manually drafting seed questions across the 52 topics, targeting concrete, culturally salient activities such as meal timing, gendered after-work routines, gift-giving customs, and burial practices. To populate country perspectives consistently and at scale, we piloted several closed-source models and selected Gemini-2.5-Flash for its quality and speed in parallel multi-perspective prompting: for each language $\times$ country pair (e.g., Arabic: Egypt, KSA, Levantine, Morocco; English: USA, UK, Australia; Spanish: Bolivia, Mexico, Spain; Portuguese: Brazil, Mozambique, Portugal; Hindi: India, Nepal), the model was instructed to act as a *country-specific cultural expert* and answer in that country's *dialect*. In the same pass we generated a standardized *Western control* answer (in MSA for Arabic) to serve as a neutral reference without introducing translation confounds.

After generation, we performed *existence filtering* to remove questions that do not apply to a given culture (e.g., asking about an ingredient never used in that region). We then transformed each item into final multiple-choice format, ensuring that each option was dialect-specific and semantically distinct; a semantic similarity pass plus manual review removed near-duplicates to guarantee discriminative answer sets. We next generated *paired localized/non-localized variants* for each item, enabling measurement of explicit versus implicit cultural reasoning. Finally, we packaged MCQ and open-ended splits, computed per-language statistics (see Table 7).

D.2 Dataset Statistics

The resulting dataset (Table 7) provides dense, balanced coverage across five languages and fourteen countries, with near-uniform counts per language–country variant ($\approx$1,372–1,607 questions per variant) and a total of 45,354 items. Localized and non-localized forms are balanced overall (57.7% vs. 42.3%), enabling clean estimation of the Localization Gap. The breadth across 52 topics (see Table 6) and depth per topic ($\approx$40–45 items) provide statistical headroom for per-topic and per-country analyses, bias detection, and mechanistic interpretability studies such as activation patching

and sparse-feature steering. In short, the combination of thematic coverage, dialectal specificity, validated cultural applicability, and contrastive pairing turns everyday cultural knowledge into a *rigorous, reproducible benchmark for evaluating and improving situated cultural adaptation of LLMs*.

Table 6: Dataset Topics by Thematic Category. The dataset spans 52 topics across 5 cultural contexts (Moroccan, Egyptian, Saudi Arabian, Levantine, and American), providing a rich lens into daily life, norms, and practices.

Category	Topics Covered	Cultural Dimensions
Daily Meals & Food Culture	• Breakfast • Lunch • Dinner • Snacks • Desserts • Fruits • Eating habits	Traditional dishes, meal timing, eating etiquette, food preferences, dietary restrictions, communal vs. individual eating styles
Daily Routines & Activities	• Before going to work/college • During commute • After work/uni (men) • After work/uni (women) • Free time activities • Household tasks	Gender-specific routines, time use, leisure preferences, division of domestic labor, work-life balance
Family & Social Relations	• Parent-child interactions • Parent-child activities • Grandparent-grandchild activities • Sibling relationships • Cousin relationships • Colleagues (work/college)	Family hierarchy, respect protocols, intergenerational differences, kinship obligations, personal vs. professional boundaries
Communication & Social Etiquette	• Greetings (verbal) • Non-verbal communication • Hospitality customs • Punctuality • Cleanliness habits	Greeting formulas, body language, guest treatment protocols, perception of time, hygiene norms and practices
Financial & Economic Practices	• Saving habits • Debt and loans • Financial discussions • Inheritance	Attitudes toward money, saving and spending strategies, debt perception, investment customs, inheritance rules
Ceremonies & Life Events	• Weddings (dowry, food, venue) • Wedding logistics and music • Gender-specific ceremonies • Burial and mourning (before, during, after)	Marriage rituals, dowry negotiations, celebration styles, gender segregation, death rituals, mourning practices
Holidays & Celebrations	• Religious holidays (before) • Religious holidays (during) • Non-religious holidays • Gift-giving customs	Religious observances, secular celebrations, festive preparations, gift exchange traditions, symbolic meaning
Cultural Expression & Recreation	• Childhood games (indoor/outdoor) • Local songs and dances • Musical instruments • Idioms and proverbs • Agricultural practices	Traditional games, folk music, dance forms, linguistic expressions, agricultural customs, community recreation

Table 7: Multilingual dataset statistics (per country and language totals).

Language	Country	Localized	Non-localized	Total
English	USA	1,372	1,372	2,744
	UK	1,372	1,372	2,744
	Australia	1,372	1,372	2,744
	Subtotal	*4,116*	*4,116*	*8,232*
Spanish	Bolivia	1,536	1,536	3,072
	Mexico	1,535	1,535	3,070
	Spain	1,536	1,536	3,072
	Subtotal	*4,607*	*4,607*	*9,214*
Portuguese	Brazil	1,607	1,607	3,214
	Mozambique	1,607	1,607	3,214
	Portugal	1,606	1,606	3,212
	Subtotal	*4,820*	*4,820*	*9,640*
Hindi	India	1,550	1,550	3,100
	Nepal	1,550	1,550	3,100
	Subtotal	*3,100*	*3,100*	*6,200*
Arabic	Egypt	1,509	1,509	3,018
	Saudi Arabia (KSA)	1,509	1,509	3,018
	Levantine	1,508	1,508	3,016
	Morocco	1,508	1,508	3,016
	Subtotal	*6,034*	*6,034*	*12,068*
Total		**22,677**	**22,677**	**45,354**