# Preference Optimization via Interventional Sparse Feature Swaps

Anonymous ACL submission

## Abstract

Preference optimization improves the alignment of large language models with human preferences, but often introduces the alignment tax by disturbing internal representations related to general capability. We propose SwapPO, a mediator-aware preference optimization framework based on sparse autoencoder (SAE) features. SwapPO first estimates feature-level alignment and capability credits through counterfactual activation interventions, and separates SAE features into alignment-private mediators and capability-sharing mediators. During training, SwapPO applies a private-mediator swap loss to guide preference updates through alignment-specific pathways, while using leakage-aware regularization to constrain excessive drift in shared mediators. Experiments on Gemma-2-2B and Gemma-2-9B show that SwapPO achieves competitive preference-alignment performance across standard benchmarks while better preserving general capabilities under the evaluated setting. Further analyses show that shared-mediator drift is more predictive of capability degradation than global policy drift, and that SwapPO induces concentrated sparse-feature changes without introducing inference-time overhead.

## 1 Introduction

In recent years, large language models have demonstrated strong capabilities in natural language understanding, text generation, code generation, and complex reasoning (Chang et al., 2024). However, since pretraining mainly relies on large-scale unsupervised corpora, model behaviors do not naturally align with human preferences (Lin et al., 2024). Therefore, how to effectively align models during post-training while preserving their original general capabilities has become a key problem in large language model research.

Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022) is a classical alignment paradigm, but it typically involves multiple stages, including reward modeling, sampling, and policy optimization, making the training process complex and computationally expensive. Direct Preference Optimization (DPO) (Rafailov et al., 2023) reparameterizes the reward function and converts preference learning into a direct classification-style optimization objective, thereby avoiding explicit reinforcement learning and significantly simplifying alignment training (Ji et al., 2026). Nevertheless, preference optimization still faces an important challenge: while improving preference alignment, the model may forget or damage the general capabilities acquired during pretraining, a phenomenon commonly referred to as the alignment tax (Lin et al., 2024). Prior studies have shown that RLHF-based alignment can induce an alignment–forgetting trade-off, where gains in human-preference alignment may come at the cost of capability preservation (Askell et al., 2021; Huang et al., 2025; Young, 2026).

Existing methods usually mitigate this problem by imposing global constraints on the policy distribution, output probabilities, or overall representation space, such as reference-model constraints, margin constraints, token-level KL regularization, or representation-level regularization (Wang et al., 2024; Rafailov et al., 2023; Meng et al., 2024; Zeng et al., 2024; Liu et al., 2025c). Recently, Feature-level constrained Preference Optimization (FPO) (Yin et al., 2025) further introduced sparse autoencoders (SAEs) into preference optimization and imposed feature-level constraints to improve training stability and controllability. However, such constraints still primarily regulate feature drift as a whole, without explicitly distinguishing the functional roles of different features in preference alignment and general capability preservation. In fact, not all feature drift necessarily leads to capability degradation: some features mainly serve preference alignment and can be safely updated, while

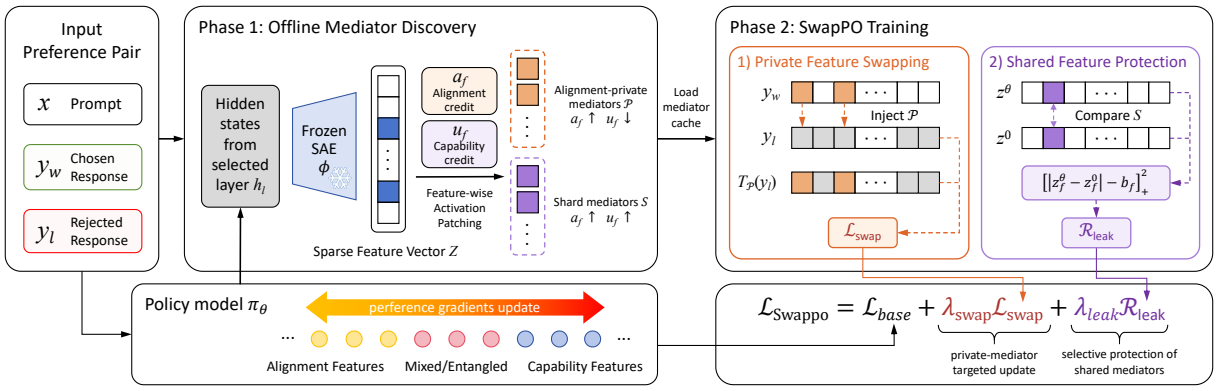


Figure 1: Mechanistic overview of SwapPO. A frozen SAE decomposes hidden states into sparse features, where counterfactual interventions identify alignment-private and capability-sharing mediators. SwapPO then performs private-mediator swaps while constraining shared-mediator drift.

other features jointly participate in both alignment and general capability, and their drift is more likely to cause capability loss.

We propose SwapPO, a preference optimization framework based on interventional sparse-feature decomposition, as shown in Figure 1. A frozen SAE first maps model hidden states into a sparse feature space, where counterfactual interventions estimate each feature's contributions to preference alignment and general capability. These credits are used to partition SAE features into alignment-private mediators and capability-sharing mediators. During training, SwapPO injects private features from chosen responses into rejected responses, guiding preference updates through alignment-specific pathways. Meanwhile, capability-sharing mediators are selectively constrained when their deviations from the reference model exceed feature-specific safe thresholds. This design enables targeted alignment updates while limiting excessive drift in capability-related mediators, providing a fine-grained mechanism for balancing alignment and capability preservation.

The main contributions are as follows:

- Feature-level interventional analysis identifies shared-mediator drift as a more informative indicator of capability degradation than global policy drift or total SAE drift.
- We propose a counterfactual intervention-based feature discovery method to identify alignment-private and shared mediators in the SAE feature space.
- SwapPO combines private-mediator feature swaps with leakage-aware regularization, improving preference alignment while limiting capability-related shared-mediator drift.

## 2 Related Work

**Preference Optimization.** LLM alignment is commonly achieved by optimizing preferences over paired responses. Early approaches rely on RLHF (Christiano et al., 2017; Lambert, 2026) with Proximal Policy Optimization (PPO) (Schulman et al., 2017; Ouyang et al., 2022), but often incur high computational cost and training instability (Casper et al., 2023). Recent work has therefore shifted toward offline preference optimization. DPO reformulates preference learning as a log-likelihood ratio objective derived from the Bradley–Terry model, enabling more stable and efficient alignment (Rafailov et al., 2023; Liu et al., 2025a). Subsequent methods further improve this paradigm: SimPO (Meng et al., 2024) and AlphaDPO (Wu et al., 2025) introduce margin-based objectives, TDPO (Zeng et al., 2024) extends preference optimization to the token level, and TGDPO (Zhu et al., 2025) and FocalPO (Liu et al., 2025b) refine training through weighting strategies. APO (D'Oosterlinck et al., 2025) and SafeDPO (Kim et al., 2025) further enhance stability with additional constraints. Despite these advances, most existing methods operate primarily at the output-probability level and provide limited control over internal representations. As a result, preference updates may affect broad regions of the representation space, entangling alignment-relevant changes with capability-related features and contributing to the alignment tax (Lin et al., 2024; Young, 2026).

**Mechanism-Aware Alignment.** Recent work has begun to approach alignment from a mechanistic perspective rather than relying solely on

output-level objectives (Bai et al., 2022; Gao et al., 2023). A growing line of research uses SAEs to decompose hidden states into sparse and interpretable features, often combined with causal interventions such as activation patching to localize components responsible for specific behaviors (Gao et al., 2024; Marks et al., 2024; Shu et al., 2025). Building on this perspective, several methods introduce feature- or representation-level constraints into alignment. For example, FPO (Yin et al., 2025) constrains activation changes in SAE feature space to reduce representation drift; REPO (Sepahvand et al., 2026) edits or erases undesirable representations; NSPO (Niu et al., 2026) projects gradients into the null space of capability-related subspaces; and OGPSA (Sun et al., 2026) enforces orthogonal gradient constraints to protect capability-relevant directions. While these methods improve global controllability, they generally do not explicitly distinguish causal mediators with different functional roles in alignment and capability preservation. In contrast, SwapPO identifies alignment-private mediators and capability-sharing mediators through counterfactual SAE interventions, and uses this decomposition to route preference updates through private mediator pathways while constraining drift in shared mediators.

# 3 Method

## 3.1 Preliminaries and Problem Setup

Let $\pi_\theta$ denote the policy model to be optimized. SwapPO uses a SimPO-style objective as the base preference loss and introduces additional feature-level constraints to regulate representation drift. Given a preference pair $(x, y_w, y_l)$, where $y_w$ is preferred over $y_l$, we define the normalized single-response score as

$$r_\theta(x, y) = \frac{\beta}{|y|} \log \pi_\theta(y|x), \tag{1}$$

where $|y|$ is the response length and $\beta$ is a scaling coefficient. The preference margin is

$$m_\theta(x, y_w, y_l) = r_\theta(x, y_w) - r_\theta(x, y_l) - \gamma, \tag{2}$$

where $\gamma$ is a fixed target margin. The base loss is

$$\mathcal{L}_{\text{base}} = -\mathbb{E}_{(x,y_w,y_l)} \log \sigma\left(m_\theta(x, y_w, y_l)\right). \tag{3}$$

This objective encourages the chosen response to obtain a higher normalized likelihood than the rejected response by a fixed margin, without using an explicit reference-policy ratio.

To analyze and constrain internal representation changes, we introduce a frozen anchor model $\pi_{\theta_0}$, where $\theta_0$ denotes the model parameters before preference optimization, and a frozen SAE $E$. The anchor model is not used in the SimPO loss; it only provides a representation baseline for measuring feature drift. Given a hidden state $h$ from a selected layer, the SAE maps it into sparse feature activations $z = E(h)$. For feature $f$, we denote its activation under the current policy as $z_f^\theta(x, y)$ and its activation under the anchor model as $z_f^0(x, y)$. All mediator credits used for constructing the mediator bank are estimated with this frozen anchor model $\pi_{\theta_0}$, while $\pi_\theta$ denotes the trainable policy optimized during SwapPO training.

SwapPO aims to improve preference alignment while selectively controlling feature-level drift. It therefore distinguishes alignment-private mediators, which can be optimized for alignment, from capability-sharing mediators, which should be protected to mitigate alignment tax.

## 3.2 Mediator Credit Estimation

To quantify the role of individual SAE features in preference alignment and general capability, SwapPO estimates feature-level causal credits through counterfactual activation interventions. For offline mediator discovery, we assume that, in a local neighborhood of the frozen discovery model, the SimPO-style response score can be approximated by a first-order expansion over SAE feature activations:

$$r_{\theta_0}(x, y) \approx \sum_f \alpha_f z_f^{\theta_0}(x, y) + \epsilon_r, \tag{4}$$

where $\alpha_f$ denotes the local sensitivity of the response score to feature $f$, and $\epsilon_r$ captures higher-order residuals. This approximation provides a local feature-level view of the preference signal and motivates estimating the causal contribution of each feature by intervening on its activation.

We first define the alignment credit, which measures how much a feature contributes to preference discrimination. For feature $f$, the alignment credit is defined as

$$a_f = \frac{1}{2}\mathbb{E}_{(x,y_w,y_l)}\Big[r_{\theta_0}(x, T_f^{y_w \to y_l}(y_l)) - r_{\theta_0}(x, y_l) + r_{\theta_0}(x, y_w) - r_{\theta_0}(x, T_f^{y_l \to y_w}(y_w))\Big]. \tag{5}$$

Here, $T_f^{y_w \to y_l}(y_l)$ denotes a counterfactual forward pass in which only feature $f$ in the representation

of $y_l$ is replaced by the corresponding activation from $y_w$. Symmetrically, $T_f^{y_l \to y_w}(y_w)$ removes the preferred-side activation by replacing it with the rejected-side activation. The first term measures whether injecting feature $f$ from the chosen response improves the rejected response score, while the second term measures whether removing the same feature from the chosen response decreases its score. Their symmetric average reduces directional bias and yields a more stable estimate of the feature's alignment contribution.

To identify features that also support general capability, we introduce a capability scoring function $g_{\theta_0}(x, c)$, instantiated as the length-normalized teacher-forced log-likelihood under the frozen discovery model. Given a capability contrast pair $(x, c_w, c_l)$, where $c_w$ is stronger than $c_l$ under the capability criterion, the capability credit of feature $f$ is defined as

$$u_f = \frac{1}{2}\mathbb{E}_{(x,c_w,c_l)}\Big[g_{\theta_0}(x, T_f^{c_w \to c_l}(c_l)) - g_{\theta_0}(x, c_l) + g_{\theta_0}(x, c_w) - g_{\theta_0}(x, T_f^{c_l \to c_w}(c_w))\Big]. \tag{6}$$

This quantity estimates whether feature $f$ lies on a pathway that is important for preserving general capability, rather than merely improving preference alignment.

We then compare the two credits using a feature-level private score:

$$p_f = \frac{|a_f|}{|u_f| + \epsilon}, \tag{7}$$

where $\epsilon$ is a small constant for numerical stability. A high value of $p_f$ indicates that feature $f$ has a strong alignment effect but limited capability interference. In contrast, features with both large $|a_f|$ and large $|u_f|$ are likely to be shared mediators between alignment and capability.

Based on these quantities, we partition SAE features into two sets. The alignment-private set $\mathcal{P}$ contains features with high private scores and sufficient alignment credit:

$$\mathcal{P} = \{f : p_f \in \mathrm{TopK}(p),\ |a_f| > \tau_a\}. \tag{8}$$

The shared set $\mathcal{S}$ contains features that contribute substantially to both preference alignment and general capability:

$$\mathcal{S} = \{f : |a_f| > \tau_a,\ |u_f| > \tau_u,\ f \notin \mathcal{P}\}. \tag{9}$$

where $\tau_a$ and $\tau_u$ are credit thresholds.

This decomposition leads to the mediator leakage hypothesis: under preference optimization, drift in shared mediators $\mathcal{S}$ is more closely associated with capability degradation than uniform drift across all SAE features. Therefore, SwapPO encourages preference updates to act through alignment-private mediators in $\mathcal{P}$ while constraining drift in capability-sharing mediators in $\mathcal{S}$.

### 3.3 Mediator Bank Construction

Before training, SwapPO performs offline mediator discovery on a calibration dataset disjoint from both training and evaluation data. For each preference or capability contrast pair, we extract hidden states from a selected target layer and encode them with the frozen SAE. Since SAE activations are sparse, interventions are restricted to active features only.

For each active feature $f$, we compute its activation difference between the two responses, decode the single-feature difference back to hidden-state space using the SAE decoder direction, and inject the resulting perturbation into the target layer through a forward hook. For preference pairs $(x, y_w, y_l)$, we recompute the SimPO-style response score $r_{\theta_0}(x, y)$ to estimate the alignment credit $a_f$. For capability contrast pairs $(x, c_w, c_l)$, we recompute the capability score $g_{\theta_0}(x, c)$ to estimate the capability credit $u_f$.

To reduce noise in credit estimation, we apply bootstrap stability filtering. The calibration set is split into multiple subsets, credits are estimated independently on each subset, and only features whose credit signs remain consistent across most subsets are retained. After filtering, we select the alignment-private set $\mathcal{P}$ by ranking retained features according to the private score $p_f$. The shared set $\mathcal{S}$ consists of retained features with both substantial alignment and capability credits, excluding those already selected into $\mathcal{P}$.

For each shared feature, we assign a leakage weight according to its capability importance:

$$w_f = \frac{|u_f|}{\sum_{j \in \mathcal{S}} |u_j| + \epsilon}, \quad f \in \mathcal{S}. \tag{10}$$

Thus, features that are more important for general capability receive stronger protection in the training objective.

### 3.4 Training Objective

After obtaining the alignment-private set $\mathcal{P}$ and the shared set $\mathcal{S}$, SwapPO optimizes a SimPO-style

base loss together with two feature-level terms: a private-mediator swap loss and a shared-mediator leakage regularization. The base loss $\mathcal{L}_{\text{base}}$ is defined in the previous section and provides the primary preference optimization signal.

For each preference pair $(x, y_w, y_l)$, we construct a counterfactual forward pass by replacing the alignment-private feature activations of $y_l$ with the corresponding activations from $y_w$. We denote this intervention as $T_{\mathcal{P}}^{y_w \to y_l}(y_l)$, where the operation is applied to the hidden representation rather than to the text sequence itself. The source activations from $y_w$ are detached during the intervention, so the loss encourages the model to respond to alignment-private mediator directions instead of trivially modifying the source features.

The swap loss is defined as

$$\mathcal{L}_{\text{swap}} = -\mathbb{E}_{(x,y_w,y_l)} \log \sigma\Big(r_\theta\big(x, T_{\mathcal{P}}^{y_w \to y_l}(y_l)\big) - r_\theta(x, y_l)\Big). \tag{11}$$

This term encourages the injected alignment-private features from the chosen response to increase the score of the rejected response. In this way, SwapPO reinforces preference-relevant directions in $\mathcal{P}$ without directly intervening on shared capability-related features.

To protect capability-relevant pathways, we apply a mediator-aware leakage regularization over the shared set $\mathcal{S}$:

$$R_{\text{leak}} = \mathbb{E}_{(x,y)} \sum_{f \in \mathcal{S}} w_f \left[\left|z_f^\theta(x,y) - z_f^0(x,y)\right| - b_f\right]_+^2, \tag{12}$$

where $z_f^\theta(x,y)$ and $z_f^0(x,y)$ denote the activations of feature $f$ under the current policy and the frozen anchor model, respectively. The coefficient $w_f$ is the leakage weight assigned according to the capability credit of feature $f$, and $b_f$ is a feature-specific tolerance threshold. This regularization penalizes only excessive drift beyond the tolerance margin, allowing harmless feature movement while constraining capability-sensitive shared mediators.

The overall SwapPO objective is

$$\mathcal{L}_{\text{SwapPO}} = \mathcal{L}_{\text{base}} + \lambda_{\text{swap}} \mathcal{L}_{\text{swap}} + \lambda_{\text{leak}} R_{\text{leak}}, \tag{13}$$

where $\lambda_{\text{swap}}$ and $\lambda_{\text{leak}}$ control the strengths of the private-mediator swap loss and the shared-mediator leakage regularization. In our experiments, the mediator bank $(\mathcal{P}, \mathcal{S}, w_f)$ is constructed once on the held-out calibration split before preference optimization and then kept fixed during training.

# 4 Experiments

## 4.1 Experimental Setup

**Models and Training.** We evaluate SwapPO on Gemma-2-2B and Gemma-2-9B (Team et al., 2024). Gemma-2-2B is used for full preference evaluation and mechanistic analysis, while Gemma-2-9B is used to test scalability. All methods are trained on UltraFeedback-binarized (Cui et al., 2024) using the same train/calibration/validation split: the training split is used for preference optimization, the calibration split for mediator discovery, and the validation split for monitoring. Experiments are run on 8 $\times$ A100 80G GPUs. SwapPO uses the SAE, mediator bank, and swap operation only during training, and therefore introduces no inference-time overhead.

**Implementation Details.** All methods are optimized with AdamW using a learning rate of $5 \times 10^{-7}$, 150 warmup steps, and a constant learning-rate schedule. The optimizer uses $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-8}$, and a weight decay of $10^{-4}$. Training runs for 1,000 steps with an effective batch size of 8. Unless otherwise specified, baseline methods use the same optimization budget and data split as SwapPO.

**Sparse Feature Configuration.** SwapPO uses frozen Gemma Scope (Lieberum et al., 2024) SAEs with a width of 16k to extract sparse features from intermediate model representations. The SAE is used only for feature encoding, mediator discovery, and sparse feature intervention during training, and is not updated during preference optimization. Mediator discovery is performed on the calibration split and completed before preference optimization.

**Evaluation Benchmarks.** Following standard preference-optimization evaluation protocols, we evaluate external preference quality on AlpacaEval 2.0 (Li et al., 2023; Dubois et al., 2024) and Arena-Hard (Li et al., 2024). For AlpacaEval 2.0, we report raw win rate (WR) and length-controlled win rate (LC). For Arena-Hard, we report WR and LC, and additionally report style-controlled win rate (SC) when available. In addition to these external benchmarks, Acc denotes pairwise accuracy on the held-out UltraFeedback preference set, measuring the fraction of examples for which the optimized policy assigns higher token-wise likelihood to the chosen response than to the rejected response. $H_{\text{2-gram}}$ denotes prompt-conditioned 2-gram Shan-

Table 1: Main results on AlpacaEval 2 and Arena-Hard, with auxiliary diagnostics. Acc and $H_{2\text{-gram}}$ are computed under the Gemma-2-2B setting. Best and second-best results within each column are marked in bold and underlined.

| Method | Gemma-2-2B | | | | | Gemma-2-9B | | | | | Acc(%) | $H_{2\text{-gram}}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AlpacaEval 2 | | Arena-Hard | | | AlpacaEval 2 | | Arena-Hard | | | | |
| | LC (%) | WR (%) | SC (%) | LC (%) | WR (%) | LC (%) | WR (%) | SC (%) | LC (%) | WR (%) | | |
| SFT | 33.0 | 31.7 | 30.0 | 31.2 | 30.8 | 49.8 | 38.4 | 33.0 | 43.0 | 42.5 | 55.8 | 5.16 |
| DPO(Rafailov et al., 2023) | **39.5** | 39.0 | 42.4 | 42.7 | 43.1 | 69.5 | 65.8 | 46.2 | 56.1 | 58.4 | 57.9 | <u>5.60</u> |
| TDPO(Zeng et al., 2024) | 38.1 | 38.4 | 42.1 | 42.8 | 43.5 | 70.8 | <u>65.9</u> | 46.4 | 57.0 | 59.2 | <u>58.2</u> | <u>5.60</u> |
| SimPO(Meng et al., 2024) | 38.4 | 38.9 | 41.9 | 42.3 | 42.6 | 70.8 | 65.4 | 46.1 | 56.4 | 58.0 | 56.4 | 5.59 |
| AlphaDPO(Wu et al., 2025) | 38.3 | 39.5 | 42.3 | 42.8 | 43.3 | 71.1 | 65.7 | <u>47.4</u> | 58.5 | 59.6 | 57.8 | 5.54 |
| FPO(Yin et al., 2025) | 39.1 | <u>40.2</u> | <u>42.7</u> | **43.1** | **43.6** | <u>71.4</u> | 65.8 | 47.0 | <u>58.6</u> | <u>60.3</u> | 58.1 | 5.52 |
| NSPO(Niu et al., 2026) | 39.3 | 39.5 | 41.2 | 40.1 | 42.9 | 71.3 | 65.0 | 46.2 | 57.8 | 59.7 | 58.0 | 5.56 |
| SwapPO | <u>39.4</u> | **40.4** | **43.2** | <u>42.9</u> | **43.6** | **71.8** | **66.1** | **47.8** | **58.8** | **60.4** | **58.8** | **6.17** |

non entropy, computed from repeated generations for each prompt to measure within-prompt generation diversity.

**Baselines.** SFT is included as the supervised fine-tuning baseline before preference optimization. The main preference-optimization baselines include DPO (Rafailov et al., 2023), SimPO (Meng et al., 2024), AlphaDPO (Wu et al., 2025), FPO (Yin et al., 2025), and NSPO (Niu et al., 2026). All methods are trained on the same preference data and evaluated under the same protocol.

### 4.2 Main Results

Table 1 reports the main results. Overall, SwapPO achieves competitive and consistently strong performance across both Gemma-2-2B and Gemma-2-9B. On Gemma-2-2B, preference-optimization methods obtain relatively close scores, suggesting that external judge-based gains are partly saturated at the smaller scale. SwapPO remains among the strongest methods, achieving the best AlpacaEval 2 WR and Arena-Hard SC scores while remaining comparable to FPO on length-controlled metrics. On Gemma-2-9B, SwapPO obtains the best score across the reported external metrics, although the absolute margins over the strongest baselines are modest. This pattern suggests that mediator-guided sparse feature intervention is a viable alternative to broader feature-level constraints for preference optimization.

Beyond win-rate metrics, SwapPO also achieves the highest held-out UltraFeedback preference accuracy and the largest prompt-conditioned 2-gram entropy. These auxiliary diagnostics suggest that the observed preference gains are not simply due to repetitive or overly conservative generations.

To assess whether preference optimization preserves general capabilities, we summarize the aggregate capability results in Table 2. Most baselines show small positive capability drops relative to the SFT baseline, whereas SwapPO obtains the highest average capability score and the lowest capability drop under the evaluated benchmarks. These results suggest that SwapPO preserves aggregate general capability while maintaining competitive preference-alignment performance. Full benchmark-level results are provided in Appendix D.

Table 2: Capability preservation summary on Gemma-2-2B. Avg. Cap. denotes average capability, and capability drop is computed relative to the SFT baseline.

| Method | Avg. Cap. (%) ↑ | Cap. Drop ↓ |
|---|---|---|
| SFT | 37.35 | – |
| DPO | 37.23 | 0.12 |
| TDPO | 37.11 | 0.24 |
| SimPO | 37.22 | 0.13 |
| AlphaDPO | 37.11 | 0.24 |
| FPO | 37.48 | -0.13 |
| NSPO | 37.17 | 0.18 |
| SwapPO | **37.55** | **-0.20** |

### 4.3 Policy Drift Analysis

We analyze global policy drift by computing post-hoc sequence-level KL divergence between each optimized policy and the frozen reference model. As shown in Figure 2, preference optimization induces policy drift on both chosen responses $y_w$ and rejected responses $y_l$. SimPO shows the largest late-stage drift, while DPO increases steadily during training. In contrast, FPO and SwapPO exhibit more controlled trajectories, with SwapPO generally remaining below DPO and SimPO. This suggests that SwapPO can improve preferences without relying on excessive global distributional shift. Since KL divergence provides only a coarse view of policy movement, we next examine whether the relevant changes are concentrated in sparse mediator features.

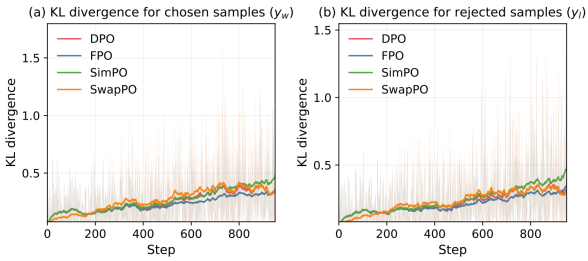


Figure 2: Global policy drift measured by sequence-level KL divergence. (a) KL divergence computed on chosen samples $y_w$. (b) KL divergence computed on rejected samples $y_l$.

### 4.4 Sparse Mediator Geometry and Feature Dynamics

To understand how SwapPO changes the model internally, we represent each SAE feature by two patch-based causal credits: alignment credit $a_f$ and capability credit $u_f$. The resulting $(u_f, a_f)$ plane gives a functional role space for sparse features, where the mediator bank is obtained by selecting a compact set of private and shared mediators from the full SAE basis. As shown in Figure 3(a), the alignment-relevant structure is highly sparse rather than broadly distributed across all SAE features, and the mediator bank occupies a compact region of the feature space.

We then compare the reference SFT checkpoint with the trained SwapPO checkpoint in the same credit space. The migration pattern in Figure 3(b) shows that post-training movement is concentrated in a small number of selected mediators, especially private mediators, while most shared mediators remain comparatively localized. To quantify whether these movements correspond to sparse alignment gain, we compute $\Delta a_f = a_f^{\text{after}} - a_f^{\text{before}}$, keep positive increments, and sort mediators by contribution. Figure 3(c) shows that the top three mediators account for $91.7\%$ of the total positive alignment-credit increase within the selected mediator bank.

These results indicate that SwapPO does not induce a diffuse update across the latent basis. Instead, it concentrates alignment-relevant functional changes into a small private-mediator subset while keeping most shared mediators comparatively stable. Consistent with this aggregate pattern, we further provide token-level case studies in Appendix G, which illustrate that private mediators mainly change around task-critical tokens while shared mediators remain more constrained.

Table 3: Mechanistic predictor analysis of capability degradation. Higher $R^2$ and Spearman correlation indicate stronger explanatory power, while lower LORO RMSE indicates better out-of-run generalization.

| Diagnostic | $R^2$ | $\rho$ | LORO RMSE |
|---|---|---|---|
| Shared-mediator drift | **0.81** | **0.88** | **0.36** |
| Total SAE drift | 0.57 | 0.71 | 0.61 |
| KL divergence | 0.42 | 0.59 | 0.72 |
| Parameter distance | 0.34 | 0.47 | 0.79 |
| Null-space energy | 0.49 | 0.65 | 0.67 |

### 4.5 Mechanistic Predictor Analysis

To further examine whether capability degradation is better explained by localized mediator drift or by coarse global movement, we regress the held-out capability-drop target against several diagnostic signals. These diagnostics include shared-mediator drift, total SAE drift, sequence-level KL divergence, parameter distance, and null-space energy. We report $R^2$, Spearman's rank correlation $\rho$, and leave-one-run-out (LORO) RMSE. As shown in Table 3, shared-mediator drift achieves the highest $R^2$, the strongest rank correlation, and the lowest LORO RMSE. This indicates that capability degradation is better explained by drift in capability-relevant sparse mediators than by coarse measures of global policy movement. In other words, alignment tax is not only related to how far the policy moves from the reference model, but also to where this movement occurs in the sparse feature space. This result supports the design of SwapPO: selectively protecting shared mediators targets the feature subset most predictive of capability loss, while leaving alignment-private mediators available for preference optimization.

### 4.6 Resource Overhead

SwapPO adds a one-time offline mediator discovery stage and a sparse feature intervention during training. The discovery stage is run before preference optimization and can be reused when the backbone model, SAE layer, and discovery configuration remain unchanged. During training, the extra resource usage mainly comes from extracting SAE activations at a single layer and applying the swap and leakage terms on the selected mediator bank. As shown in Figure 4, the cached-reference version of SwapPO reaches 53.56 GiB peak GPU memory under the matched Gemma-2-2B single-GPU setting. Thus, SwapPO remains in the same practical memory regime as feature-level baselines,

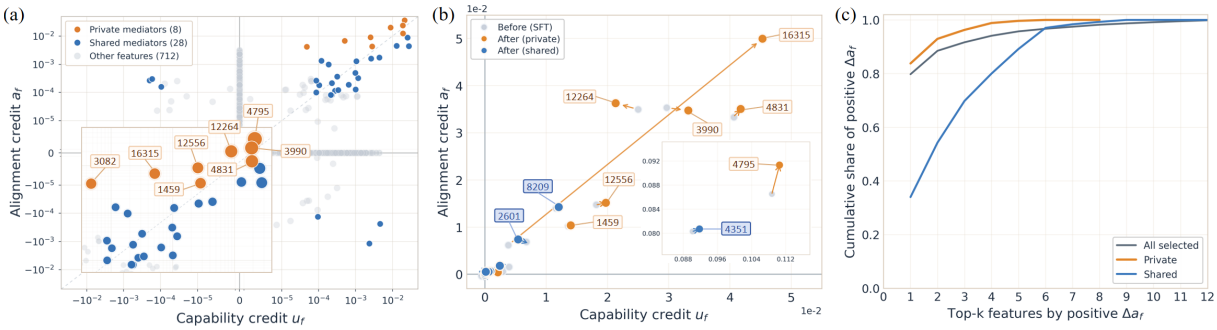


Figure 3: Sparse feature geometry and concentrated private-mediator editing. (a) Static distribution of SAE features in the capability/alignment credit plane. Alignment-relevant structure is sparse, and the discovered mediator bank occupies a compact region of the feature space. (b) Checkpoint-level mediator migration from the reference SFT model to the trained SwapPO model. Functional movement is concentrated in a small number of selected mediators. (c) Cumulative concentration of positive alignment-credit increase. The sharp curve shows that most post-training alignment gain is explained by only a few mediators.

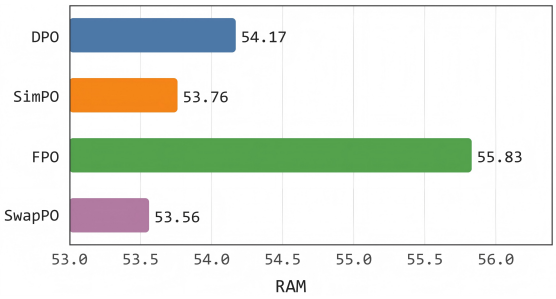


Figure 4: Peak GPU memory usage during preference optimization on Gemma-2-2B.

Table 4: Ablation study of SwapPO components on Gemma-2-2B.

| Variant vs. Full | Acc(%) | WR (%) | LC(%) |
|---|---|---|---|
| **Full** | **58.8** | – | – |
| Swap-only | 57.5 | 47.5 | 47.0 |
| Leak-only | 58.5 | 49.1 | 49.7 |
| Random mediators | 58.2 | 48.5 | 47.4 |
| Top-activation mediators | 58.3 | 47.5 | 47.2 |
| All-feature leak | 57.6 | 48.7 | 48.2 |
| No private/shared split | 57.4 | 47.7 | 47.6 |

while adding no inference-time overhead because the SAE and mediator bank are used only during training.

### 4.7 Ablation Studies

Each ablated variant is directly compared against the full SwapPO model on AlpacaEval 2, and a values below 50 indicate degradation relative to the full method. Table 4 shows that removing either private-mediator swapping or shared-mediator leakage control degrades performance. Swap-only performs worse than Leak-only, suggesting that shared-mediator protection is especially important. Random mediators and Top-activation mediators also underperform the full model, indicating that causal mediator discovery is more effective than random or activation-based feature selection. In addition, All-feature leak and No private/shared split both reduce performance, showing that uniform feature constraints and removing the private/shared decomposition weaken SwapPO. Overall, these results confirm that SwapPO benefits from combining private-mediator routing with selective shared-mediator protection.

## 5 Conclusion

This paper introduced SwapPO, a mediator-aware preference optimization framework for mitigating alignment tax through sparse feature-level control. SwapPO identifies alignment-private and capability-sharing SAE mediators via counterfactual activation interventions, then combines private-mediator swapping with leakage-aware regularization to guide alignment updates while limiting unnecessary shared-feature drift. Experiments on Gemma-2-2B and Gemma-2-9B show that SwapPO achieves strong preference-alignment performance across standard benchmarks without introducing inference-time overhead.

The results further support a feature-level view of alignment tax: capability degradation is more closely associated with drift in capability-sharing mediators than with global policy or representation shift. This suggests that future preference optimization methods should move beyond coarse global constraints and instead regulate functionally decomposed internal feature pathways.

## Limitations

SwapPO relies on the quality and stability of the SAE used for mediator discovery. If the SAE dictionary contains noisy, unstable, or poorly interpretable features, the resulting private/shared mediator partition may be imperfect. The method also depends on the selected layer and the representativeness of the held-out calibration data used for activation patching, so mediator credits estimated on one model or data distribution may not fully transfer to other settings. Although SwapPO does not introduce inference-time overhead, it requires an additional offline mediator discovery stage before preference optimization. Our experiments are mainly conducted on English preference-alignment benchmarks and Gemma-family models, so further validation is needed across more languages, model families, modalities, and safety-critical deployment scenarios. Finally, the proposed mediator leakage analysis should be viewed as evidence for a feature-level explanation of alignment tax rather than a complete causal theory of all capability degradation after alignment tuning.

## References


Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, and 1 others. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

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. *arXiv preprint arXiv:2204.05862*.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, and 1 others. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

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

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2024. Ultrafeedback: Boosting language models with high-quality feedback.

Karel D'Oosterlinck, Winnie Xu, Chris Develder, Thomas Demeester, Amanpreet Singh, Christopher Potts, Douwe Kiela, and Shikib Mehri. 2025. Anchored preference optimization and contrastive revisions: Addressing underspecification in alignment. *Transactions of the Association for Computational Linguistics*, 13:442–460.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2024. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*.

Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, Zachary Yahn, Yichang Xu, and Ling Liu. 2025. Safety tax: Safety alignment makes your large reasoning models less reasonable. *arXiv preprint arXiv:2503.00555*.

Miaomiao Ji, Yanqiu Wu, Zhibin Wu, Shoujin Wang, Jian Yang, Mark Dras, and Usman Naseem. 2026. A survey of progress in llm alignment from the perspective of reward design. *IEEE Transactions on Artificial Intelligence*.

Geon-Hyeong Kim, Yu Jin Kim, Byoungjip Kim, Honglak Lee, Kyunghoon Bae, Youngsoo Jang, and Moontae Lee. 2025. Safedpo: A simple approach to direct preference optimization with enhanced safety. *arXiv preprint arXiv:2505.20065*.

Nathan Lambert. 2026. Reinforcement learning from human feedback. *Preprint*, arXiv:2504.12501.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse

autoencoders everywhere all at once on gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 278–300.

Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, and 1 others. 2024. Mitigating the alignment tax of rlhf. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 580–606.

Shunyu Liu, Wenkai Fang, Zetian Hu, Junjie Zhang, Yang Zhou, Kongcheng Zhang, Rongcheng Tu, Ting-En Lin, Fei Huang, Mingli Song, and 1 others. 2025a. A survey of direct preference optimization. *arXiv preprint arXiv:2503.11701*.

Tong Liu, Xiao Yu, Wenxuan Zhou, Jindong Gu, and Volker Tresp. 2025b. Focalpo: Enhancing preference optimizing by focusing on correct preference rankings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 256–267.

Yixin Liu, Pengfei Liu, and Arman Cohan. 2025c. Understanding reference policies in direct preference optimization. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8022–8037.

Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2024. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *arXiv preprint arXiv:2403.19647*.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235.

Yifan Niu, Han Xiao, Dongyi Liu, Nuo Chen, and Jia Li. 2026. Mitigating the safety alignment tax with null-space constrained policy optimization. *arXiv preprint arXiv:2512.11391*.

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.

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.

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

Nazanin Mohammadi Sepahvand, Eleni Triantafillou, Hugo Larochelle, Doina Precup, Daniel M Roy, and Gintare Karolina Dziugaite. 2026. Detoxifying llms via representation erasure-based preference optimization. *arXiv preprint arXiv:2602.23391*.

Dong Shu, Xuansheng Wu, Haiyan Zhao, Daking Rai, Ziyu Yao, Ninghao Liu, and Mengnan Du. 2025. A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models. *arXiv preprint arXiv:2503.05613*.

Guanglong Sun, Siyuan Zhang, Liyuan Wang, Jun Zhu, Hang Su, and Yi Zhong. 2026. Safety alignment as continual learning: Mitigating the alignment tax via orthogonal gradient projection. *arXiv preprint arXiv:2602.07892*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Zhichao Wang, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Xiang-Bo Mao, Sitaram Asur, and 1 others. 2024. A comprehensive survey of llm alignment techniques: Rlhf, rlaif, ppo, dpo and more. *arXiv preprint arXiv:2407.16216*.

Junkang Wu, Xue Wang, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. 2025. AlphaDPO: Adaptive reward margin for direct preference optimization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 67793–67809. PMLR.

Qingyu Yin, Chak Tou Leong, Hongbo Zhang, Minjun Zhu, Hanqi Yan, Qiang Zhang, Yulan He, Wenjie Li, Jun Wang, Yue Zhang, and 1 others. 2025. Constrain alignment with sparse autoencoders. In *International Conference on Machine Learning*, pages 72349–72363. PMLR.

Robin Young. 2026. What is the alignment tax? *arXiv preprint arXiv:2603.00047*.

Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. 2024. Token-level direct preference optimization. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Mingkang Zhu, Xi Chen, Zhongdao Wang, Bei Yu, Hengshuang Zhao, and Jiaya Jia. 2025. TGDPO: Harnessing token-level reward guidance for enhancing direct preference optimization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 79727–79745. PMLR.

# Appendix

## A Details of Capability Credit Estimation

The capability credit $u_f$ is estimated on a held-out capability calibration set $\mathcal{D}_{cap}$, which is disjoint from the preference-training set and all final evaluation benchmarks. Each example is represented as a capability contrast pair $(x, c^+, c^-)$, where $c^+$ denotes a task-correct or capability-preserving response and $c^-$ denotes an incorrect or lower-quality response under the same prompt. In all experiments, the capability scoring function $g_{\theta_0}(x, c)$ is instantiated as the length-normalized teacher-forced log-likelihood under the frozen discovery model:

$$g_\theta(x, c) = \frac{1}{|c|} \sum_{t=1}^{|c|} \log \pi_{\theta_0}(c_t \mid x, c_{<t}). \qquad (14)$$

This score is used only for offline mediator discovery. It is not included in the online preference optimization objective and does not introduce an additional reward model.

For each SAE feature $f$, we estimate its capability credit by symmetric single-feature activation patching:

$$u_f = \frac{1}{2}\mathbb{E}_{(x,c^+,c^-)\sim\mathcal{D}_{\text{cap}}}\Big[g_{\theta_0}(x, T_f^{c^+\to c^-}(c^-)) - g_{\theta_0}(x, c^-) + g_{\theta_0}(x, c^+) - g_{\theta_0}(x, T_f^{c^-\to c^+}(c^+))\Big]. \qquad (15)$$

The first difference measures whether injecting feature $f$ from the stronger capability response into the weaker response improves its score. The second difference measures whether replacing the same feature in the stronger response with the weaker-side activation decreases its score. Their symmetric average reduces directional bias and yields a more stable estimate of whether feature $f$ supports general capability.

The estimated $u_f$ is used only to distinguish alignment-private mediators from capability-sharing mediators. Features with high alignment credit $a_f$ but low capability credit $u_f$ are assigned to the private mediator set, while features with jointly high $a_f$ and $u_f$ are treated as shared mediators and receive leakage-aware protection during training. Because $g_{\theta_0}$ is evaluated under the frozen discovery model, $u_f$ is fixed after mediator discovery and is not updated during online preference optimization.

## B Hyperparameter Settings

Table 5 summarizes the training hyperparameters used in our main experiments. Since SwapPO additionally requires an offline mediator discovery stage, we report the corresponding discovery, intervention, and scaling configurations in Table 6.

### B.1 Mediator Discovery Configuration

Mediator discovery is performed once on the held-out calibration split before preference optimization, and the resulting mediator bank is kept fixed during training. All credit estimation, patching, and drift measurements are response-masked: prompt tokens and padding tokens are excluded through the label mask. When chosen and rejected responses have different effective lengths, padded or non-response positions receive zero weight, and private-target terms are computed only on overlapping response positions.

For the Gemma-2-9B scaling study, we do not reuse the 2B mediator bank. We repeat mediator discovery from scratch using the corresponding Gemma-2-9B final-layer SAE, because SAE features are model- and layer-specific. The resulting 9B mediator bank contains 8 private mediators and 507 shared mediators. Figure 5 visualizes the resulting 9B feature-credit geometry.

### B.2 Parameter Sensitivity of SwapPO

We further analyze the sensitivity of SwapPO to its sparse intervention strength and regularization strategy. Table 7 shows a clear trend: overly weak sparse interventions lead to limited alignment gains, whereas moderately stronger interventions consistently improve both pairwise accuracy and reward margin. However, aggressive updates also tend to increase shared drift, indicating a tradeoff between alignment improvement and representation stability.

The results further show that private-target regularization plays a central role in obtaining a favorable tradeoff. In particular, cosine-based private-target regularization yields substantially better behavior than weaker or pointwise alternatives, improving both alignment accuracy and stability. We also observe that a more conservative intervention regime, with slightly smaller edit scale, fewer edited positions, and more selective feature retention, achieves the best balance between reward margin and shared-feature preservation. Overall, these findings suggest that SwapPO benefits from suffi-

Table 5: Hyperparameters of the formal preference optimization experiments.

| Method | Key Hyperparameters |
|---|---|
| DPO | $\beta = 0.1$ |
| TDPO | $\beta = 0.1,\ \alpha = 0.5$ |
| SimPO | $\beta = 2.0,\ \gamma = 0.5,\ \alpha = 1.0$ |
| AlphaDPO | $\beta = 2.0,\ \alpha = 0.3,\ \gamma = 0.5$ |
| FPO | $\beta = 0.1,\ \alpha = 0.1,\ \gamma = 0.1,\ \text{use_mse} = \text{true}$ |
| NSPO | $\beta = 0.1,\ \alpha = 10^{-4},\ n_{\text{cal_batches}} = 4$ |
| SwapPO | $\beta = 0.1,\ K = 8,\ \alpha = 1.0,\ \gamma = 0.5,\ \lambda_{\text{swap}} = 0.0011,\ \lambda_{\text{leak}} = 0.045,\ \lambda_{\text{private}} = 0.02$ |

Table 6: Mediator discovery, intervention, and scaling configuration. The Gemma-2-2B setting is used for the main SwapPO experiments. The Gemma-2-9B setting is used for the scaling diagnostic.

| Item | Gemma-2-2B main | Gemma-2-9B scaling diagnostic |
|---|---|---|
| Backbone | Gemma-2-2B | Gemma-2-9B |
| SAE source | gemma-scope-2b-pt-res | gemma-scope-9b-pt-res |
| SAE layer | Layer 25 | Layer 41 |
| SAE width | 16,384 features | 16,384 features |
| Alignment credit samples | 500 preference pairs | 500 preference pairs |
| Capability credit samples | 200 contrast pairs | 200 contrast pairs |
| Discovery patch positions | top 8 response-token positions | top 8 response-token positions |
| Bootstrap rounds | 5 | 5 |
| Private score $\epsilon$ | 0.01 | 0.01 |
| Shared mediators discovered | 28 features | 507 features |
| Shared barrier $b_f$ | 0.9 quantile of shared activation | 0.9 quantile of shared activation |
| Barrier scale | 1.0 | 1.0 |
| Training swap positions | top 24 rejected response-token positions | top 24 rejected response-token positions |
| Training private feature count $K$ | 8 | 8 |

ciently strong but carefully controlled sparse interventions, rather than from either overly weak or overly aggressive updates.

## C Prompt-conditioned 2-gram Shannon Entropy.

To evaluate whether preference optimization preserves the diversity of the model's conditional generation distribution, we adopt a prompt-conditioned 2-gram Shannon entropy as an auxiliary diversity metric. Unlike standard benchmark metrics that evaluate response quality or preference win rate, this metric focuses on whether the model still produces diverse local expressions when repeatedly sampled under the same input prompt. This is particularly relevant for preference optimization, since an aligned model may obtain a higher win rate by converging to a narrower and more conservative response pattern.

For each evaluation prompt $x_i$, we independently sample $K = 8$ responses from the same model using a fixed decoding configuration, such as the same temperature, top-$p$, maximum generation length, and random seed protocol. The sampled responses for prompt $x_i$ are denoted as

$$\mathcal{Y}_i = \left\{ y_i^{(1)}, y_i^{(2)}, \ldots, y_i^{(K)} \right\}. \tag{16}$$

Here, $y_i^{(k)}$ represents the $k$-th independently sampled response for the same prompt $x_i$. Keeping the decoding configuration fixed across all methods ensures that the measured diversity difference mainly comes from the optimized model itself rather than from sampling hyperparameters.

Before computing the entropy, we remove system prompts, chat templates, special tokens, and other formatting artifacts that are not part of the actual generated response. Each cleaned response is then tokenized, and all adjacent 2-grams are extracted. For a response $y_i^{(k)}$ with token sequence

$$y_i^{(k)} = \left( w_{i,1}^{(k)}, w_{i,2}^{(k)}, \ldots, w_{i,T_{i,k}}^{(k)} \right), \tag{17}$$

its 2-gram sequence is defined as

$$\mathcal{G}_i^{(k)} = \left\{ \left( w_{i,t}^{(k)}, w_{i,t+1}^{(k)} \right) \mid 1 \le t < T_{i,k} \right\}. \tag{18}$$

We then aggregate the 2-grams from all $K$ sampled responses of the same prompt into a prompt-level

Table 7: Hyperparameter sensitivity of SwapPO on held-out internal validation. All settings use the same mediator bank and $K = 8$. Acc denotes held-out pairwise accuracy. Margin denotes the average chosen–rejected log-probability gap, and Shared drift measures perturbation magnitude on shared mediators. Higher Acc and Margin are better, while lower Shared drift is better.

| $\lambda_{\text{swap}}$ | $\lambda_{\text{leak}}$ | $\lambda_{\text{priv}}$ | Private-target loss | Scale | Max pos. | Norm ratio | Quantile | Acc (%)↑ | Margin ↑ | Shared drift ↓ |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.0005 | 0.020 | 0.000 | – | 0.050 | 16 | 0.15 | 0.50 | 49.5 | 0.00054 | 0.807 |
| 0.0010 | 0.030 | 0.000 | – | 0.075 | 16 | 0.20 | 0.50 | 49.5 | 0.00506 | 0.721 |
| 0.0015 | 0.030 | 0.000 | – | 0.100 | 24 | 0.25 | 0.25 | 52.0 | 0.00856 | 0.837 |
| 0.0010 | 0.050 | 0.000 | – | 0.050 | 24 | 0.15 | 0.75 | 51.0 | 0.00322 | 0.700 |
| 0.0012 | 0.050 | 0.002 | SmoothL1 | 0.075 | 24 | 0.20 | 0.50 | 51.5 | 0.00194 | 0.718 |
| 0.0012 | 0.050 | 0.005 | SmoothL1 | 0.075 | 24 | 0.20 | 0.50 | 56.5 | 0.00963 | 0.789 |
| 0.0012 | 0.050 | 0.020 | Cosine | 0.075 | 24 | 0.20 | 0.50 | **59.5** | 0.01470 | 0.548 |
| 0.0011 | 0.045 | 0.020 | Cosine | 0.070 | 20 | 0.18 | 0.60 | 58.8 | **0.01549** | **0.504** |

multiset:

$$\mathcal{G}_i = \bigcup_{k=1}^{K} \mathcal{G}_i^{(k)}. \tag{19}$$

The use of a multiset means that repeated 2-grams are counted multiple times. Therefore, the metric captures not only how many different local expressions appear, but also how concentrated or dispersed their empirical distribution is.

Let $c_i(g)$ denote the number of occurrences of a 2-gram $g$ in $\mathcal{G}_i$. The empirical probability of $g$ under prompt $x_i$ is computed as

$$p_i(g) = \frac{c_i(g)}{\sum_{g'} c_i(g')}. \tag{20}$$

The summation in the denominator counts the total number of extracted 2-grams across all $K$ sampled responses for prompt $x_i$. Based on this empirical distribution, the prompt-level 2-gram Shannon entropy is defined as

$$H_i^{(2)} = -\sum_{g \in \mathrm{supp}(\mathcal{G}_i)} p_i(g) \log p_i(g), \tag{21}$$

where $\mathrm{supp}(\mathcal{G}_i)$ denotes the set of distinct 2-grams observed in the sampled responses for prompt $x_i$. We use the natural logarithm in all experiments.

Finally, the overall diversity score is obtained by averaging the prompt-level entropy over all $N$ evaluation prompts:

$$H^{(2)} = \frac{1}{N} \sum_{i=1}^{N} H_i^{(2)}. \tag{22}$$

This averaging strategy makes the metric prompt-conditioned: entropy is first computed within the response distribution of each individual prompt, and only then averaged across the benchmark. As a result, $H^{(2)}$ reflects whether the model maintains diverse responses under the same input condition, rather than merely measuring the lexical coverage of the entire benchmark.

This distinction is important. A corpus-level entropy computed by concatenating one response per prompt can be strongly affected by dataset size, topic coverage, average response length, and prompt heterogeneity. In contrast, the proposed prompt-conditioned entropy directly measures local generation diversity under repeated sampling. If a model repeatedly generates nearly identical responses for the same prompt, its 2-gram distribution becomes highly concentrated, leading to a lower $H_i^{(2)}$. Conversely, if the model produces semantically reasonable but locally varied expressions, the empirical 2-gram distribution becomes more dispersed, resulting in a higher entropy value.

In our analysis, $H^{(2)}$ is not used as a standalone indicator of response quality. A higher entropy does not necessarily imply better alignment, since noisy or incoherent generations may also increase the diversity of local 2-grams. Instead, this metric is used as an auxiliary diagnostic together with win-rate, length-controlled win-rate, and capability-oriented evaluations. When a method improves preference performance while maintaining a comparable $H^{(2)}$, it suggests that the method improves alignment without substantially collapsing the conditional generation distribution. In contrast, a large drop in $H^{(2)}$ may indicate that the optimization process narrows the model toward more repetitive or conservative response patterns.

## D Mitigating Alignment Tax

Table 8 evaluates capability preservation after preference optimization. We use the SFT model as the reference baseline and report five capability-oriented benchmarks. The average capability score summarizes performance across these benchmarks, while capability drop measures the difference

Table 8: Capability preservation after preference optimization. We use the SFT model as the reference baseline, and capability drop is computed relative to this baseline. Gray cells indicate scores below the SFT baseline. Lower capability drop indicates better preservation of general capabilities.

| Method | MMLU ↑ | GSM8K ↑ | HumanEval ↑ | ARC-C ↑ | TruthfulQA ↑ | Avg. Cap. ↑ | Cap. Drop ↓ |
|---|---|---|---|---|---|---|---|
| SFT | 52.85 | 25.25 | 19.51 | 53.75 | 35.40 | 37.35 | – |
| DPO | 52.86 | 25.25 | 19.51 | 53.50 | 35.05 | 37.23 | 0.12 |
| TDPO | 52.89 | 25.25 | 18.90 | 53.50 | 34.99 | 37.11 | 0.24 |
| SimPO | 52.81 | 25.25 | 19.51 | 53.50 | 35.03 | 37.22 | 0.13 |
| AlphaDPO | 52.86 | 25.32 | 18.90 | 53.41 | 35.04 | 37.11 | 0.24 |
| FPO | 52.88 | 25.55 | **20.12** | **53.84** | 35.02 | 37.48 | -0.13 |
| NSPO | **53.01** | 25.55 | 18.90 | 53.33 | 35.05 | 37.17 | 0.18 |
| SwapPO | 52.93 | **25.63** | **20.12** | **53.84** | **35.25** | **37.55** | **-0.20** |

from the SFT baseline. Overall, most preference-optimization baselines show small positive capability drops, indicating mild degradation on general-purpose tasks after alignment tuning. In contrast, SwapPO achieves negative capability drops, suggesting that its aggregate capability scores remain slightly above the SFT baseline.

Among all preference-optimized policies, SwapPO obtains the highest average capability score and the lowest capability drop. It matches the best scores on HumanEval and ARC-Challenge, achieves the strongest TruthfulQA score among preference-optimized methods, and remains competitive on MMLU and GSM8K. These results indicate that SwapPO does not obtain alignment gains by sacrificing broad model capability. Instead, by separating preference-relevant private mediators from shared capability-related mediators, SwapPO better preserves general-purpose behavior, suggesting that it reduces observable alignment tax under the evaluated setting.

## E KL Divergence Computation

For SwapPO, the KL divergence is used only as a post-hoc diagnostic metric to measure the global policy drift from the frozen reference model. It is not an explicit training loss term.

Given a frozen reference model $\pi_{\text{ref}}$ and a SwapPO policy checkpoint $\pi_\theta$, we compute the sequence-level KL approximation on the validation preference pairs. For each prompt-response pair $(x, y)$, where $y$ can be either the chosen response $y_w$ or the rejected response $y_l$, we concatenate the prompt and response as the model input, but compute log probabilities only on the response tokens. Prompt tokens and padding tokens are masked out and do not contribute to the calculation.

For a response $y = (y_1, \ldots, y_{|y|})$, the KL value used in our analysis is computed as

$$D_{\text{SeqKL}}(x, y; \pi_{\text{ref}} \| \pi_\theta) = \frac{1}{|y|} \sum_{t=1}^{|y|} \Big[ \log \pi_{\text{ref}}(y_t \mid x, y_{<t}) - \log \pi_\theta(y_t \mid x, y_{<t}) \Big]. \tag{23}$$

The normalization by $|y|$ gives the average token-level log-ratio and avoids making longer responses dominate the KL value.

For a validation set

$$\mathcal{D}_{\text{val}} = \{(x_i, y_{w,i}, y_{l,i})\}_{i=1}^{N}, \tag{24}$$

we compute the chosen-response KL of SwapPO as

$$D_{\text{KL}}^{w} = \frac{1}{N} \sum_{i=1}^{N} D_{\text{SeqKL}}(x_i, y_{w,i}; \pi_{\text{ref}} \| \pi_\theta), \tag{25}$$

and the rejected-response KL as

$$D_{\text{KL}}^{l} = \frac{1}{N} \sum_{i=1}^{N} D_{\text{SeqKL}}(x_i, y_{l,i}; \pi_{\text{ref}} \| \pi_\theta). \tag{26}$$

These two quantities are computed for each saved SwapPO checkpoint, producing the KL curves over training steps.

In implementation, the reference model is loaded once and kept frozen, while each SwapPO checkpoint is loaded as the current policy model. For each validation sample, we first compute the average response-token log probability under $\pi_{\text{ref}}$, then compute the same quantity under $\pi_\theta$. The per-sample KL approximation is simply their difference:

$$\widehat{D}_{\text{SeqKL}}(x, y) = \overline{\log \pi_{\text{ref}}(y \mid x)} - \overline{\log \pi_\theta(y \mid x)}. \tag{27}$$

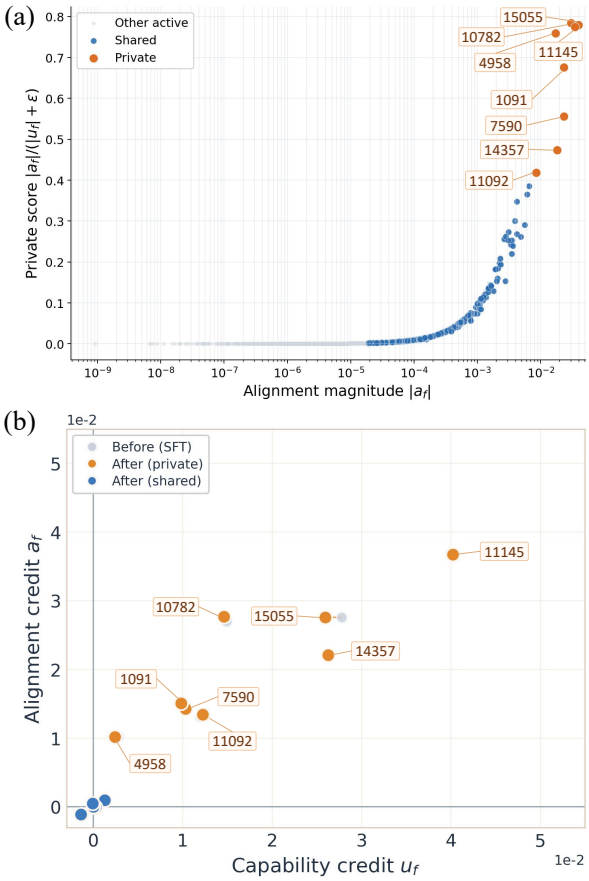


Figure 5: Gemma-2-9B mediator selectivity and feature-credit migration. (a) The discovered private mediators occupy the high-private-score region of the final-layer SAE feature space, separating them from shared and other active features. (b) After SwapPO training, feature-credit movement is concentrated on selected private mediators, while shared mediators remain comparatively stable.

The final reported value is the mean over all validation samples.

This metric reflects how far the SwapPO policy moves away from the frozen reference model during training. A larger value indicates stronger global policy drift, while a smaller value indicates a more conservative update. In our analysis, SwapPO shows a controlled non-monotonic pattern: the KL divergence first increases in the early stage, reflecting the necessary policy adjustment for preference learning, and then decreases or stabilizes as training proceeds. This suggests that SwapPO does not continuously accumulate global distributional drift, but instead guides optimization toward more localized mediator-level updates.

# F Mechanistic Predictor Analysis

We compute mechanistic predictors on saved checkpoints from all compared preference-optimization methods. Checkpoints are collected every 50 training steps from each method, resulting in 20 checkpoints per method. For each checkpoint, the capability-drop target is computed as

$$\Delta_{\text{cap}}(\theta) = \text{AvgCap}(\theta_{\text{SFT}}) - \text{AvgCap}(\theta), \quad (28)$$

where $\text{AvgCap}(\cdot)$ denotes the average score over the capability-oriented benchmarks used in Table 2.

For each checkpoint, we compute the following diagnostic signals.

**Shared-mediator drift.** Shared-mediator drift measures the weighted activation deviation of capability-sharing mediators from the frozen anchor model:

$$D_{\mathcal{S}}(\theta) = \mathbb{E}_{(x,y)} \sum_{f \in \mathcal{S}} w_f \left| z_f^{\theta}(x,y) - z_f^{\theta_0}(x,y) \right|. \quad (29)$$

Here, $\mathcal{S}$ denotes the shared-mediator set, $w_f$ is the leakage weight derived from capability credit, and $\theta_0$ is the frozen anchor model.

**Total SAE drift.** Total SAE drift measures the average feature-space deviation over all SAE features:

$$D_{\text{SAE}}(\theta) = \mathbb{E}_{(x,y)} \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} \left| z_f^{\theta}(x,y) - z_f^{\theta_0}(x,y) \right|, \quad (30)$$

where $\mathcal{F}$ is the full SAE feature set.

**KL divergence.** KL divergence measures global policy drift from the frozen reference model. We compute the same response-token sequence-level KL diagnostic as described in Appendix E, and use the checkpoint-level average as the predictor.

**Parameter distance.** Parameter distance is computed as the normalized Euclidean distance between the optimized policy and the frozen anchor model:

$$D_{\text{param}}(\theta) = \frac{\|\theta - \theta_0\|_2}{\|\theta_0\|_2}. \quad (31)$$

**Null-space energy.** Null-space energy measures how much of the parameter update lies outside the estimated capability-relevant subspace. Let $P_{\mathcal{C}}$ denote the projection matrix onto the capability-relevant subspace. We compute

$$D_{\text{null}}(\theta) = \frac{\|(I - P_{\mathcal{C}})(\theta - \theta_0)\|_2^2}{\|\theta - \theta_0\|_2^2}. \quad (32)$$

## G Token-level Mediator Case Studies

To complement the aggregate sparse-feature analysis in the main text, we provide qualitative token-level visualizations on three representative preference pairs from the Gemma-2-2B experiments. As shown in Figure 7, each heatmap corresponds to one preference pair, where the horizontal axis denotes tokens in the preferred response and the vertical axis denotes selected mediator features. The color indicates the activation change of each feature after SwapPO training: red denotes increased activation and blue denotes decreased activation. Rows marked with P correspond to alignment-private mediators, while the remaining rows correspond to capability-sharing mediators.

Overall, these qualitative examples (Example 1-3) show a consistent feature-level pattern across different task types. Alignment-private mediators tend to change near task-critical decision points, including factual entities, entailment conclusions, uncertainty judgments, confidence statements, and answer-final positions. Capability-sharing mediators show more localized and constrained changes. These visualizations do not serve as standalone evidence of performance improvement, but they support the interpretation that SwapPO routes preference-relevant updates through private mediator pathways while limiting broad drift in shared capability-related features.

## H Qualitative Response Comparison

The qualitative examples (Cases 1–3) are drawn from the Gemma-2-2B experiments and illustrate where SwapPO improves response behavior beyond the aggregate metrics. The goal is not to show a wholesale rewrite of the answer style, but to show that SwapPO makes more appropriate local decisions when the response must satisfy the user's intent.

Across the three cases, the baselines exhibit different failure modes. In the Shabbat case, DPO and FPO drift away from the requested list of measures and instead generate unsupported background information about an organization. SwapPO restores the requested response type by giving concrete measures for home and synagogue observance. In the slingshot case, all models answer the question, but DPO remains generic and FPO suggests harder targets too early. SwapPO organizes the advice into a more usable progression, from simple stationary targets to harder practice settings. In the proof case, DPO repeats the goal without completing the argument, while FPO contains the core proof idea in a less clean generated format. SwapPO presents the reasoning as a complete proof by rewriting $k^2 + 2k + 1$ as $(k + 1)^2$, applying the given non-negativity fact, and stating the conclusion.

The token-level heatmaps provide a complementary mechanism view for these behavioral differences (Figure 7). Each heatmap visualizes the mediator activation delta between SwapPO and DPO on the displayed SwapPO response tokens, where red cells indicate higher activation under SwapPO and blue cells indicate lower activation. The largest changes appear around content-bearing tokens, such as Shabbat activities, target-practice instructions, and the quadratic rewrite. This suggests that SwapPO does not change the response uniformly; instead, it modifies preference-relevant features in localized regions of the answer where the useful behavior is expressed.

Taken together, these cases suggest that SwapPO's benefit is not simply verbosity or a stronger surface template. The improvements occur where the answer needs to match the requested format, turn generic advice into an executable procedure, or supply a missing reasoning step. This qualitative evidence is consistent with the intended mechanism of SwapPO: targeted preference refinement that improves task-faithful behavior while largely preserving the general response structure.

**Example 1: Factual Question Answering**

**Question:** Where and when will the play-off games of the upcoming season of the Star Sports Pro Kabaddi League be held?

**Preferred response:** The play-off games of the upcoming season of the Star Sports Pro Kabaddi League will be held in New Delhi between March 4 and 6.

**Rejected response:** The rejected response also mentions New Delhi and the relevant dates, but it is more verbose and introduces additional schedule-related explanations and irrelevant follow-up content. As a result, the answer boundary is less clean.

**Interpretation:** The preferred response directly extracts the two key pieces of information required by the question: the location, New Delhi, and the time, March 4–6. In the corresponding heatmap, alignment-private mediators show noticeable changes around factual tokens and the answer-ending region. This suggests that SwapPO does not uniformly modify all tokens, but instead concentrates preference-relevant changes near positions that determine factual grounding, answer conciseness, and response closure.

**Example 2: Positive Entailment**

**Question:** Can we conclude from "A group of athletes in blue short-and-shirt uniforms huddle together on a field" that "The athletes are on a field"?

**Preferred response:** The group of athletes is described as being on a field in the given statement. The answer is: yes. Confidence: 95%.

**Rejected response:** The rejected response drifts away from the given premise and hypothesis, discussing an unrelated scene involving other people and objects rather than making the required entailment judgment.

**Interpretation:** The correct answer is "yes" because the premise explicitly states that the athletes are on a field. In the heatmap, alignment-private mediators change around the rationale, the final "yes" decision, and the confidence statement. In contrast, capability-sharing mediators remain comparatively constrained and are not broadly amplified across the full response. This pattern suggests that SwapPO reinforces task-relevant local representations rather than simply encouraging a fixed answer format or diffuse activation increase.

**Example 3: Uncertain Entailment**

**Question:** Premise: "A slender man holds a squirming child on a city street." Can we conclude that the hypothesis "The man is the child's father" is true?

**Preferred response:** Let's solve this gradually. The premise does not provide enough information about the relationship between the man and the child. Answer is it is not possible to tell. Confidence: 95%.

**Rejected response:** The rejected response infers that the man and the child may have a strong bond and therefore that the man is likely to be the child's father. This is an unsupported inference, since the premise only states that the man holds the child.

**Interpretation:** The preferred response correctly recognizes the boundary of the evidence: the premise supports that a man is holding a child, but it does not establish a father-child relationship. In the heatmap, alignment-private mediators show clear changes around the relationship phrase, the "not possible to tell" decision, and the final answer region. This suggests that SwapPO can strengthen representations associated with evidence-bounded reasoning, where the model should avoid over-inference when the premise is insufficient.

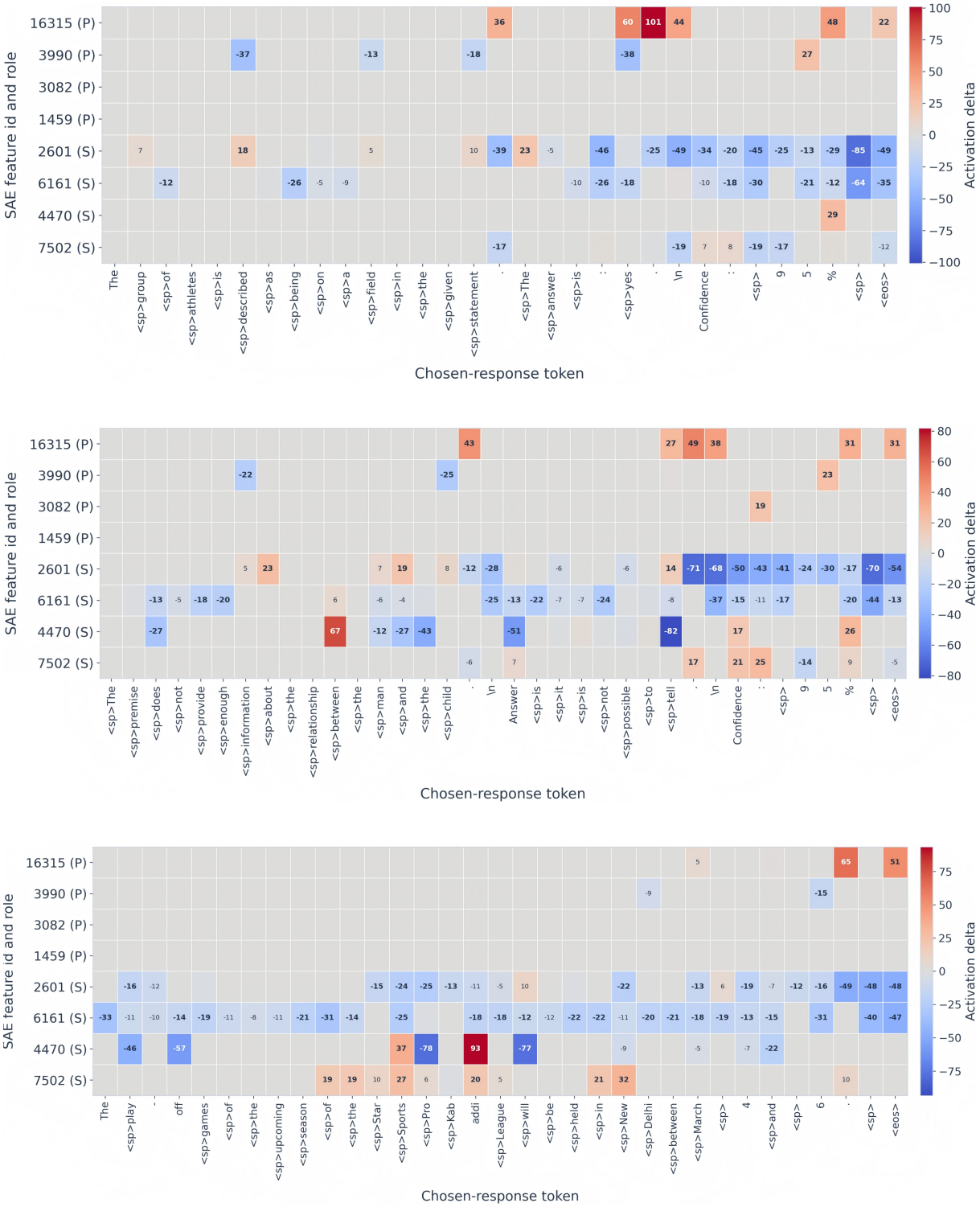


Figure 6: Token-level activation changes of selected SAE mediators on representative preference pairs. Columns correspond to tokens in the preferred response, and rows correspond to discovered mediator features. Rows marked with P denote alignment-private mediators, while the remaining rows denote capability-sharing mediators. Red indicates increased activation and blue indicates decreased activation after SwapPO training. The three examples cover factual question answering, positive entailment, and uncertain entailment.

**Case 1: Shabbat Celebration and Observance**

*Prompt:* Write a list of measures and ideas how the Sophia Jewish can foster more Shabbat celebration and observance at home and in synagogue.

**DPO response**

The Sophia Jewish is a Jewish community center in the heart of the city of Sophia, Bulgaria. The center is located in the heart of the city and is a hub for Jewish life in the city. The center offers a variety of programs and services to the Jewish community, including a synagogue, a Jewish school, a Jewish library, and a Jewish museum. The center also offers a variety of programs and services to the general public, including a Jewish cultural center, a Jewish museum, and a Jewish library.

**FPO response**

The Sophia Jewish is a Jewish community center in the heart of the city of Sophia, Bulgaria. The center is located in the heart of the city and is a hub for Jewish life in the city. The center offers a variety of programs and services to the Jewish community, including a synagogue, a Jewish school, a Jewish library, and a Jewish museum. The center also offers a variety of programs and services to the general public, including a Jewish cultural center, a Jewish museum, and a Jewish library.

**SwapPO response**

The Sophia Jewish can foster more Shabbat celebration and observance at home and in synagogue by:
1. Encouraging members to attend Shabbat services and participate in the community.
2. Providing resources and support for members to create a Shabbat atmosphere at home.
3. Offering opportunities for members to learn about and celebrate Shabbat together.
4. Promoting Shabbat as a time for rest and relaxation, and encouraging members to take time to enjoy each other's company.
5. Providing opportunities for members to connect with each other and build relationships through Shabbat activities.
6. Encouraging members to share their Shabbat.

*Interpretation:* DPO and FPO drift away from the user request. Instead of providing measures for increasing Shabbat celebration and observance, both models hallucinate unsupported institutional details about "the Sophia Jewish" as a community center in Sophia, Bulgaria. SwapPO stays closer to the instruction and gives concrete, actionable measures for both home and synagogue settings.

**Case 2: Slingshot Aiming**

*Prompt:* How do I get better at aiming with my slingshot?

**DPO response**

You can improve your aim by practicing with a slingshot. Start by aiming at a target that is close to you, such as a wall or a tree. As you get better, you can move the target further away. You can also try different types of slingshots, such as a recurve or a wrist sling.

**FPO response**

You can try to aim at a target that is far away from you. This will help you to get used to the slingshot and to improve your aim. You can also try to aim at a target that is moving, as this will help you to improve your reaction time.

**SwapPO response**

You can improve your aim by practicing with a slingshot. Start by practicing with a simple target, such as a piece of paper or a cardboard box. Aim for the center of the target and try to hit it as many times as possible. As you get better, you can try more challenging targets, such as a can or a bottle. You can also try different types of slingshots, such as a recurve or a wrist sling, to see which one works best for you.

*Interpretation:* DPO is generic and FPO jumps to far-away or moving targets too early. SwapPO gives a clearer progression: start with a simple target, practice center hits, then increase difficulty.

**Case 3: Quadratic Non-Negativity Proof**

*Prompt:* Based on the facts that have been provided, prove the following statement. Statement: "For every integer $k$, $k^2 + 2k + 1 \geq 0$." Fact: The square of any real number is non-negative.

**DPO response**

Let $k$ be an integer. We need to prove that $k^2 + 2k + 1 \geq 0$.
Fact: The square of any real number is non-negative.
The generation repeats the statement and does not perform the algebraic rewrite needed for the proof.

**FPO response**

Let $k$ be an integer. We need to show that $k^2 + 2k + 1 \geq 0$.
We can rewrite the expression as $(k + 1)^2$. Since the square of any real number is non-negative, we have $(k + 1)^2 \geq 0$. Therefore, for every integer $k$, $k^2 + 2k + 1 \geq 0$.
The content is mostly correct, but the generated answer is fragmented by repeated assistant-turn markers.

**SwapPO response**

Yes, this is true.
Proof: Let $k$ be an arbitrary integer. Then, $k^2 + 2k + 1 = (k + 1)^2$. Since the square of any real number is non-negative, we have that $(k + 1)^2 \geq 0$. Therefore, $k^2 + 2k + 1 \geq 0$.
QED.

*Interpretation:* DPO leaves the proof incomplete, and FPO gives the right reasoning in a less clean generated format. SwapPO gives the complete argument in a single coherent proof: rewrite the quadratic as a square, apply the provided fact, and conclude the inequality.

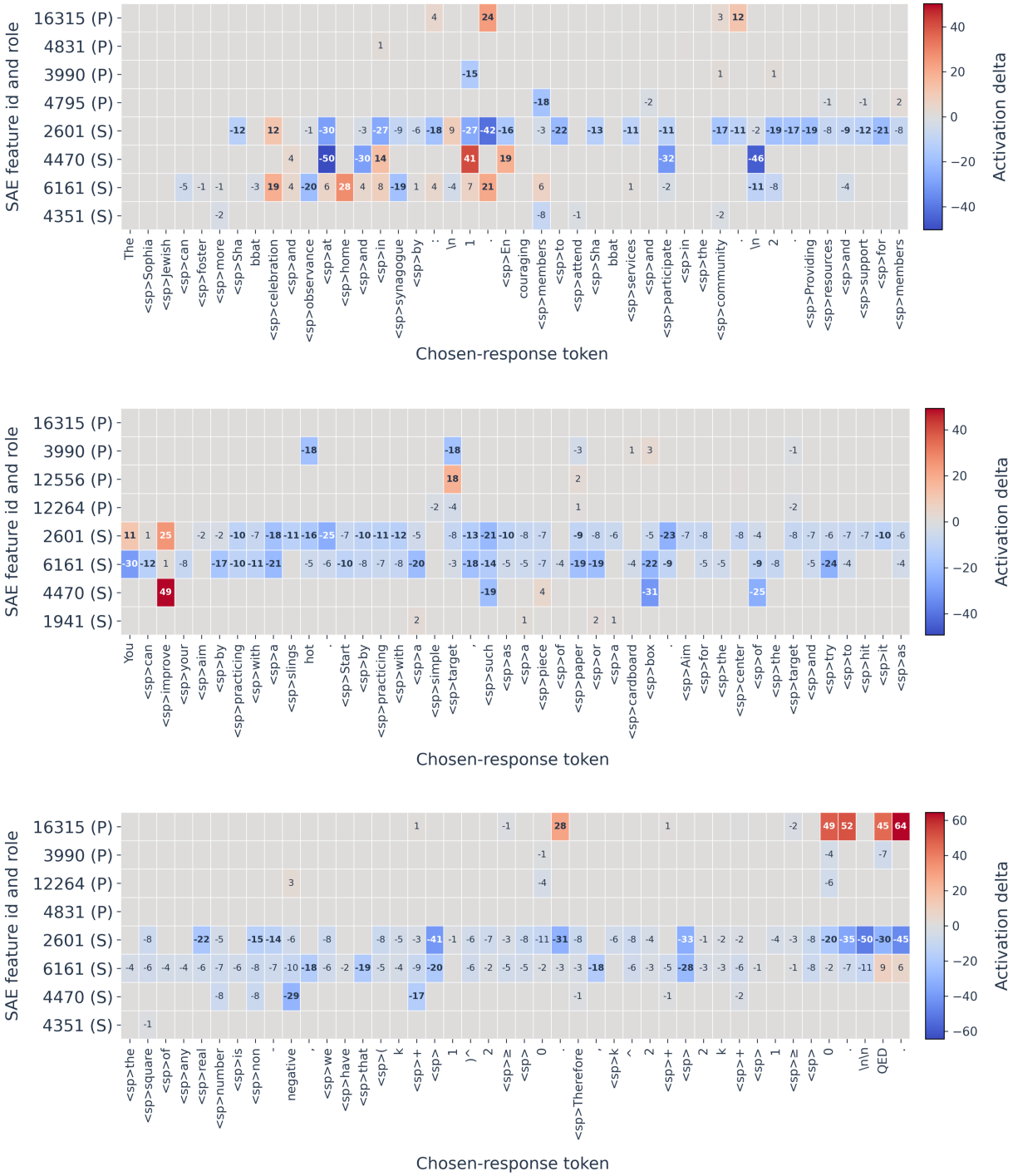


Figure 7: Token-level mediator activation changes in qualitative examples. Each heatmap shows the activation delta between SwapPO and DPO on the displayed SwapPO response tokens. Red cells indicate higher mediator activation under SwapPO, while blue cells indicate lower activation. The highlighted changes concentrate around task-relevant tokens, such as Shabbat action items, slingshot practice steps, and the quadratic proof rewrite.