MASKING FREQUENT TOKENS SHARPENS DIRECT PREFERENCE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Direct Preference Optimization (DPO) aligns language models by optimizing over sequence-level sums of token-wise implicit reward differences. However, we identify a pervasive pathology in this formulation: a disproportionately small subset of high-frequency token types dominates cumulative sequence scores while appearing symmetrically across both preferred and dispreferred responses. Specifically, under canonical Qwen tokenization on Anthropic HH-RLHF, merely 69 token types account for 55.1% of all response tokens and 85.9% of within-pair shared token mass, exhibiting substantially lower preference-side specificity than the remaining vocabulary. This symmetric ubiquity induces gradient entanglement and dilutes the discriminative preference signal propagated through the objective. To resolve this issue, we introduce *Anisotropic DPO* (ADPO) and its canonical realization, *Frequency-Hard DPO*. Using a fixed, label-agnostic vocabulary mask, our method zeroes the implicit reward contribution of high-frequency response tokens while assigning unit weight to informative positions, thereby suppressing gradient interference without modifying preference pairs, discarding context, or introducing learned parameters. Here, *anisotropy* designates non-uniform token-level objective weighting rather than representational geometry. Extensive empirical evaluations on AlpacaEval, MT-Bench, and Arena-Hard demonstrate that Frequency-Hard DPO consistently outperforms standard DPO across Qwen-2.5-7B-Instruct and Llama-3-8B-Instruct, establishing that selectively masking shared high-frequency tokens offers an effective, zero-overhead mechanism for robust preference alignment.

1 INTRODUCTION

Learning from human preferences is a standard approach to aligning language models (Ouyang et al., 2022). Direct Preference Optimization (DPO) implements this supervision through a margin between chosen and rejected responses (Rafailov et al., 2023). Each response is scored by summing token-level policy–reference log-ratios with the same explicit coefficient. Equal coefficients do not imply equal gradients: token contributions depend on their prefixes and the model's shared parameters. The question is which of these contributions should directly determine the preference update.

One difficulty is *gradient entanglement*. A margin can improve while the probabilities of both responses move in the same direction, because each likelihood change depends on the other response's gradient (Yuan et al., 2025). Related work documents declining chosen-response likelihood and displacement toward unintended alternatives (Pal et al., 2024; Razin et al., 2025). Shared response content is a natural place to investigate this coupling. Frequent token types occupy much of the occurrence mass on both sides of preference pairs, yet show weaker aggregate chosen–rejected distinction than retained types in our audits. This motivates testing whether suppressing their direct score contributions can reduce the alignment component of gradient entanglement.

We propose *Anisotropic DPO (ADPO)*, with its canonical form *Frequency-hard DPO*. ADPO estimates response-token frequencies once from both sides of training pairs and excludes the highest-frequency types from the preference score. The mask is fixed, label-agnostic, and shared by chosen and rejected responses. All tokens remain in the conditioning context and the full softmax vocabulary. There is no learned selector, auxiliary reward model, or change to the architecture. The

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

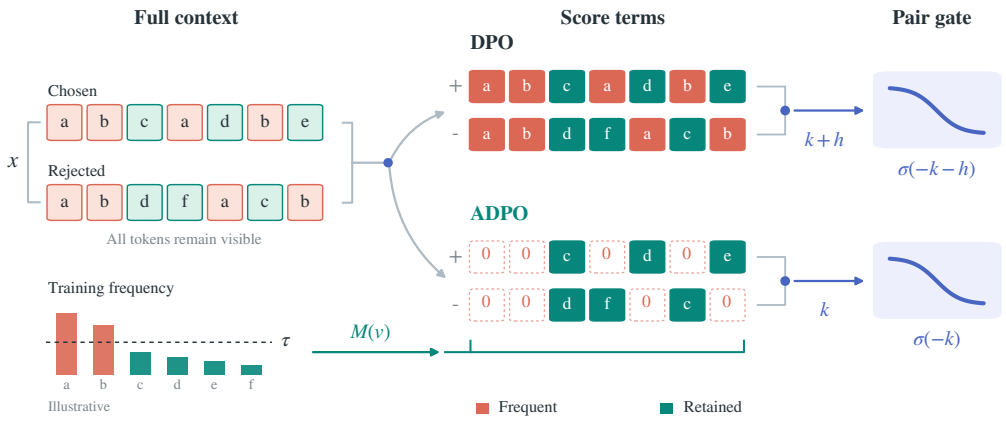

Figure 1: **ADPO masks score terms, not context.** Coral marks frequent types; teal marks retained types. Letters identify token types, not score values. Frequencies and coefficients are illustrative, not a shared positional mask. Each response keeps its full context and softmax; only coral score terms are zeroed. With beta-scaled retained and frequent margins k and h, selection changes the sigmoid gate from $\sigma(-k-h)$ to $\sigma(-k)$. Zero direct score weight does not eliminate indirect gradients through shared parameters.

intervention is *selective regularization* of preference fitting: it changes both the direct gradient terms and the pair-level sigmoid multiplier on those that remain (Figure 1).

Token reweighting itself is established prior work, including oracle-based, sparse, and policy-confidence-based approaches (Yang et al., 2025; Christopoulou et al., 2024; Yoon et al., 2025). Our focus is the simpler corpus-frequency prior. We use *anisotropic preference optimization* to describe unequal coefficients in the token-score basis, by highlighting that frequent tokens in themselves don't carry any preference signal.

This work formalizes a novel preference tuning objective by first analyzing chosen-rejected gradient cosine similarities and the preference signals of high-frequency tokens. We isolate genuine predictive information from abundant shared probability mass through structural audits and frozen reward-model comparisons, utilizing random-mask and full-likelihood controls to test the generalizability of our gradient interpretation. Our primary aim is to establish a transparent and computationally inexpensive regularizer for preference fitting. Although we separate the resulting reduction in score-gradient alignment from a general solution to representation entanglement, this targeted restriction consistently yields superior preference optimization.

Our contributions are as follows:

1. **Analysis of frequent-token preference signal.** We show that frequent types occupy substantial shared response mass yet exhibit weaker token-identity preference signal in the studied datasets.
2. **Analysis of gradient coupling.** We characterize how ADPO changes both score-gradient selection and the shared sigmoid gate, and demonstrate reduced chosen–rejected score-gradient alignment at fixed checkpoints.
3. **Frequency-based preference optimization.** We propose Anisotropic DPO (ADPO), which uses a fixed, training-derived frequency mask to suppress direct preference-score contributions from high-frequency token types while preserving the complete context and full softmax, without a learned selector or auxiliary reward model.

2 Related Work

Preference learning and direct alignment. RLHF learns reward proxies from comparisons and optimizes a policy against them (Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022). DPO replaces the explicit reward-fitting and RL stages with a policy–reference preference loss (Rafailov

et al., 2023). SLiC-HF calibrates sequence likelihoods using feedback (Zhao et al., 2023); IPO develops an alternative preference-learning objective (Azar et al., 2024). ORPO and SimPO remove the separate reference-model requirement (Hong et al., 2024; Meng et al., 2024), while KTO uses binary desirability feedback and a prospect-theoretic utility (Ethayarajh et al., 2023). SPPO uses a self-play formulation of preference learning (Wu et al., 2025). ADPO retains DPO's frozen reference and logistic pairwise loss, changing which realized token terms enter the margin rather than introducing a new preference model.

Gradient coupling and likelihood displacement. DPO-Positive addresses declining chosen-response likelihood, particularly for pairs with small edit distance (Pal et al., 2024). Yuan et al. (2025) formalize *gradient entanglement*: favorable likelihood changes depend on the cross-gradient inner product relative to both squared gradient norms. Their proposed mitigations already include pairwise gradient normalization and sparsity-regularized contextual token masking. Separately, Razin et al. (2025) characterize likelihood displacement through centered hidden-embedding similarity and use it for data filtering. AlphaPO changes reward shape (Gupta et al., 2025); normalized-reward objectives regularize the combined length-normalized probability of the preference pair and examine outlier-token likelihood changes (Im et al., 2026). ADPO instead tests a fixed frequency selector.

Token-level objectives and credit assignment. DPO has a token-level inverse-Q-learning interpretation and can perform implicit credit assignment (Rafailov et al., 2023); equal explicit score coefficients should not be confused with equal token influence. SimPO instead uses the average token log-probability of a response as its implicit reward, yielding a length-normalized, reference-free preference objective without an explicit token-level weighting or reference-model term SimPO (Meng et al., 2024). TDPO adds token-level forward-KL control (Zeng et al., 2024). TGDPO introduces token-level reward guidance (Zhu et al., 2025). These methods modify the treatment of contextual token contributions. Selective preference optimization further estimates token-level reward using a DPO-trained oracle and optimizes only selected tokens (Yang et al., 2025), while SparsePO learns sparse token-level masks to control reward and KL contributions (Christopoulou et al., 2024). These methods modify the treatment of contextual token contributions. ADPO uses no token-reward estimator or importance sampling correction, and its hard mask deliberately changes the objective rather than claiming an unbiased estimate of the original DPO gradient.

Frequency priors, nuisance signals, and evaluation. Frequency subsampling in word2vec offers a precedent for distinguishing occurrence count from task usefulness (Mikolov et al., 2013), but does not establish a preference-learning guarantee. Length-aware DPO and length-controlled evaluation show why apparent preference gains can depend on response length (Park et al., 2024; Dubois et al., 2024). Reward-model overoptimization further separates proxy fitting from the desired evaluation target (Gao et al., 2023). Our analyses accordingly distinguish occurrence mass, token-identity specificity, contextual prediction, and reward-model sensitivity.

Representation anisotropy is a distinct question. Representation degeneration (Gao et al., 2019), anisotropy from self-attention (Godey et al., 2024), and output-matrix features for text embeddings (Wu et al., 2026) concern learned representation geometry. ADPO instead defines anisotropy through unequal token-score coefficients.

3 Analysis

3.1 The Isotropic Scaling Assumption in DPO

Consider a prompt x and a response $y = (y_1, \ldots, y_L)$. For each response token y_t, define the policy-reference log-ratio as

$$a_t(\theta) = \log \pi_\theta \left(y_t \mid x, y_{<t}\right) - \log \pi_{\text{ref}} \left(y_t \mid x, y_{<t}\right). \tag{1}$$

For a chosen response y^w and a rejected response y^l, standard DPO constructs the preference margin

$$m(\theta) = \beta \left[\sum_{t=1}^{L_w} a_t^w(\theta) - \sum_{t=1}^{L_l} a_t^l(\theta) \right], \tag{2}$$

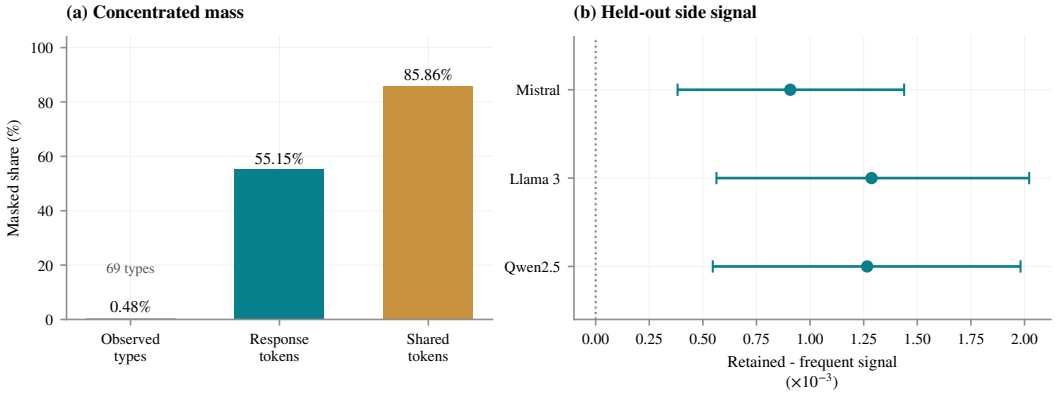

Figure 2: **Frequency-mask coverage and held-out side signal.** (a) We measure the frequent-type mask's coverage of observed vocabulary types, response occurrences, and shared chosen–rejected occurrences in HH calibration data. (b) We compare retained and frequent types using held-out preference-side associations estimated on separate training pairs. Error bars show paired confidence intervals. The panels use distinct populations.

and minimizes

$$\mathcal{L}_{\mathrm{DPO}} = -\log \sigma\left(m(\theta)\right). \tag{3}$$

Differentiating the loss with respect to the preference margin gives

$$\frac{\partial \mathcal{L}_{\mathrm{DPO}}}{\partial m} = -\sigma(-m). \tag{4}$$

Applying the chain rule, the policy gradient is therefore

$$\nabla_\theta \mathcal{L}_{\mathrm{DPO}} = -\sigma(-m)\beta \left[\sum_{t=1}^{L_w} \nabla_\theta \log \pi_\theta \left(y_t^w \mid x, y_{<t}^w\right) - \sum_{t=1}^{L_l} \nabla_\theta \log \pi_\theta \left(y_t^l \mid x, y_{<t}^l\right) \right]. \tag{5}$$

The scalar β is therefore applied identically to every realized token. This is the isotropic scaling assumption in DPO: although different tokens may produce very different gradients through the network, their contributions to the preference objective are all multiplied by the same global temperature.

3.2 A Structural Frequency Prior

Frequent types form a small vocabulary subset but account for much of the overlap between chosen and rejected responses. In most of the textual preference datasets like HH-RLHF (Bai et al., 2022), their shared occurrence mass is disproportionately large, while their token-identity association with the preference side is weaker than that of retained types (Figure 2(b) and Figure 3(b)). This motivates a fixed frequency prior for restricting direct preference fitting without removing the shared context.

We freeze each tokenizer's frequent-type set on training data, then measure its coverage in disjoint held-out responses from the same Llama-3-UltraFeedback release (Cui et al., 2023; Meng et al., 2024). These types remain common in held-out responses and account for much of their chosen–rejected overlap. Separately, Figure 2(b) measures whether token associations with the chosen or rejected side repeat on held-out pairs. Associations estimated on separate training pairs reproduce more strongly for retained types across the tokenizers examined.

3.3 Effect of Frequency Masking on Gradient Alignment

Gradient entanglement couples chosen and rejected likelihood changes (Yuan et al., 2025). We study the effect of frequency masking on chosen–rejected gradient alignment as a diagnostic for understanding why the frequency-hard objective may behave differently from standard DPO. We note that we do not intend to claim that frequency masking provides a general mechanism for eliminating

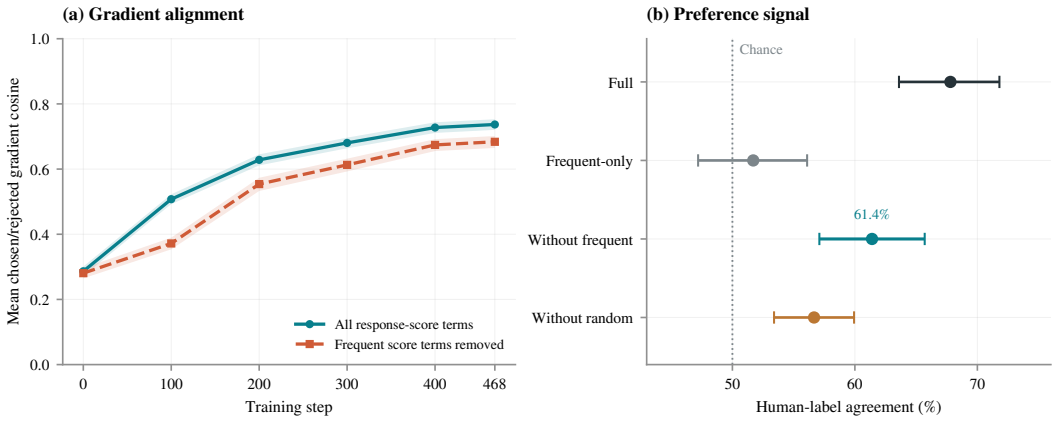

Figure 3: **Gradient alignment and preference signal.** (a) We compute mean chosen–rejected cosine for full and frequency-hard response-score gradients at saved DPO checkpoints, using identical weights, response pairs, and input context.(b) We score complete HH responses, frequent-only fragments, and frequent- or count-matched random-deletion variants with frozen PairRM, then compare rankings with the original human labels. Error bars show prompt-bootstrap intervals, and the vertical line marks chance. The panels use different samples and interventions: score masking versus response editing.

gradient entanglement. At each saved DPO checkpoint, we compare full and frequency-hard score gradients, for $s \in \{w, l\}$:

$$g_s = \nabla_\theta \sum_t \log \pi_\theta(y_t^s \mid x, y_{<t}^s),$$
$$a_s = \nabla_\theta \sum_t M(y_t^s) \log \pi_\theta(y_t^s \mid x, y_{<t}^s). \tag{6}$$

g_s uses all score terms; a_s retains unmasked terms. Both differentiate all parameters with unchanged context.

Figure 3(a) uses identical weights and response pairs within each comparison. Removing frequent score terms lowers mean cosine at every measured checkpoint, suggesting reduced chosen–rejected directional coupling under frequency-hard scoring throughout the sampled trajectory. This observation is consistent with frequency masking changing which token-level score contributions participate in the gradient, providing a diagnostic for a possible mechanism underlying the empirical behavior of the frequency-hard objective.

3.4 Preference Signal in Frequent Tokens

The frequency-hard mask is motivated by the hypothesis that highly frequent tokens carry less preference-relevant information than the tokens it retains. We therefore evaluate the preference signal associated with frequent-token content using both a frozen reward model and human preference labels.

On a held-out HH sample, we construct four response conditions: the complete response, a frequent-only fragment containing the masked token types, the complement obtained by removing frequent tokens, and a count-matched random deletion control. We score all conditions using frozen PairRM (Jiang et al., 2023) and compare the resulting rankings against the original human preference labels. As shown in Figure 3(b) and Table 1, the frequent-only fragments produce a near-zero reward margin (0.08779) and near-chance agreement with human labels (51.7%). In contrast, removing the frequent tokens retains a substantially larger reward margin (0.97335) and higher human-label agreement (61.4%).

The gap between complete and edited responses can naturally reflect the loss of response length, context, and coherence introduced by the editing procedure, which limits what can be inferred from

Table 1: **Human-label agreement and reward margins.** We evaluate the response conditions in Figure 3(b) against the original HH preference labels. Ties receive half credit; margins are order-symmetrized raw PairRM logits, not reward probabilities.

Condition	Agreement (%)	Margin
Full	67.8	1.89969
Frequent-only	51.7	0.08779
Without frequent	61.4	0.97335
Without random	56.67	0.60148

the absolute performance of the fragment conditions. Our central observation is instead that the non-frequent complement retains substantially more preference signal than the frequent-token fragment. The count-matched random deletion condition provides a supplementary control for the effect of removing a comparable number of tokens.

Taken together, these results suggest that measurable preference signal is concentrated more strongly in the non-frequent content than in the frequent-token fragments, providing empirical motivation for restricting direct DPO fitting on the frequent token types while retaining them in the causal context.

4 ANISOTROPIC DPO

Anisotropic DPO (ADPO) which is also the *Frequency-hard DPO* name the same single method: fixed binary weighting of realized token-score terms using empirical response frequencies.

4.1 THE DPO TOKEN SCORE

For preference pair i, let y_i^w and y_i^l denote the chosen and rejected responses. With the reference policy fixed, define

$$\begin{aligned} a_{it}^s(\theta) = \; & \log \pi_\theta(y_{it}^s \mid x_i, y_{i,<t}^s) \\ & - \log \pi_{\mathrm{ref}}(y_{it}^s \mid x_i, y_{i,<t}^s), \quad s \in \{w, l\}. \end{aligned} \tag{7}$$

The ordinary DPO margin and objective are

$$m_i(\theta) = \beta \left[\sum_{t=1}^{L_i^w} a_{it}^w(\theta) - \sum_{t=1}^{L_i^l} a_{it}^l(\theta) \right], \tag{8}$$

$$\mathcal{L}_{\mathrm{DPO}}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \log \sigma(m_i(\theta)). \tag{9}$$

Our intervention changes the terms included in this score, while retaining the logistic pairwise loss and frozen reference.

4.2 THE FIXED FREQUENCY MASK

For each tokenizer, count response positions in a training-calibration subset $\mathcal{D}_f$ of M pairs, using the training tokenization, truncation, and EOS handling. The count and normalized frequency of type v are

$$c(v) = \sum_{i=1}^{M} \left[c_i^w(v) + c_i^l(v) \right], \qquad \widehat{f}(v) = \frac{c(v)}{\sum_{u \in \mathcal{V}} c(u)}. \tag{10}$$

Both sides enter symmetrically, making selection invariant to swapping the preference labels. The canonical configuration uses the first $M = 4{,}000$ training pairs, the model's chat template, prompt/sequence caps of 256/512 tokens, and response EOS. Counts are recomputed for each tokenizer.

For threshold τ, define the vocabulary-level weight

$$w_\tau(v) = \mathbf{1}[\widehat{f}(v) \le \tau]. \tag{11}$$

We use the strict threshold $\tau = 0.002$, retain unseen types at unit weight, and fix w_τ throughout training. This is the mask M in the analysis. Write $\mathcal{H}_\tau = \{v : \widehat{f}(v) > \tau\}$ and $\mathcal{K}_\tau = \mathcal{V} \setminus \mathcal{H}_\tau$ for the masked and retained vocabulary sets. We have provided an ablation on τ later.

4.3 The Frequency-Hard Objective

Using Equation 7, the selected sequence score is

$$R_{\theta,\tau}(x_i, y_i^s) = \sum_{t=1}^{L_i^s} w_\tau(y_{it}^s) a_{it}^s(\theta). \tag{12}$$

Frequency-hard DPO minimizes

$$\begin{aligned} m_{i,\tau}^{\mathrm{FH}}(\theta) &= \beta \left[R_{\theta,\tau}(x_i, y_i^w) - R_{\theta,\tau}(x_i, y_i^l) \right], \\ \mathcal{L}_{\mathrm{FH}}(\theta) &= -\frac{1}{N} \sum_{i=1}^{N} \log \sigma\big(m_{i,\tau}^{\mathrm{FH}}(\theta)\big). \end{aligned} \tag{13}$$

The effective coefficient is $\beta_{it} = \beta\, w_\tau(y_{it})$. With no masked realized types, the objective reduces exactly to standard DPO.

Scores are not renormalized by retained token count. A response with no retained positions has zero direct score; if both sides have none, the pair has loss $\log 2$ and zero direct gradient, and is not dropped. All original tokens remain in the causal context and log-probabilities use the full softmax vocabulary. Zero weight is not a frozen output row or an attention mask: shared parameters can still receive indirect gradients.

Writing the full beta-scaled margin as retained k plus frequent h, ADPO changes the negative loss gradient from $\sigma(-k-h)\nabla(k+h)$ to $\sigma(-k)\nabla k$. It therefore combines score-gradient selection with pair-dependent gate reweighting.

Table 2: Token types removed by the frequency-hard mask at $\tau = 0.002$. The mask contains 50 vocabulary types whose empirical frequency exceeds the threshold.

Token type	**Tokens masked**
Function words / stopwords	`the`, `is`, `The`, `and`, `for`, `to`, `that`, `a`, `with`, `it`, `of`, `or`, `are`, `in`, `can`, `be`, `on`, `by`, `as`
Punctuation	`,`, `.`, `:`, `)`, `"`
Markdown / code markup	`*`, `**`, `**:`, `**`, `` ` ``, ` `` `
Whitespace / Newline composites	*space*, `\n`, *three-space indent*, `\t`, `.\n`, `.\n\n`, `:\n\n`, `**\n\n`

5 Results

5.1 Experimental Setup

We evaluate the proposed framework on Llama-3-8B-Instruct, and Qwen2.5-7B-Instruct. For each model, the policy is initialized from the corresponding instruction-tuned checkpoint, while an identical frozen copy is used as the reference model (wherever applicable). No supervised fine-tuning stage is performed before preference optimization.

To construct the preference data, we follow the on-policy data generation recipe introduced by SimPO. Starting from the UltraFeedback prompts, we regenerate candidate responses using the corresponding instruction-tuned model rather than using the original preference responses. For each prompt, we sample five responses from the model with a temperature of 0.8 and use RLHFlow/ArmoRM-Llama3-8B-v0.1 (Wang et al., 2024) to rank them. The highest-scoring response is selected as the chosen response y^w, while the lowest-scoring response is selected as the rejected response y^l. We generate the preference data in a single pass, following the SimPO procedure, thereby making the preference distribution closer to that of the policy being optimized.

Table 3: Main benchmark results for Qwen2.5-7B-Instruct. MT-Bench and AlpacaEval 2.0 use GPT-4-turbo as the API judge, while Arena-Hard uses GPT-4.1-mini as the API judge. $\pm$ values denote respective uncertainity estimates.

Method	FsFairX RM Mean Reward	MT-Bench Score	AlpacaEval 2.0 LC Win Rate	Arena-Hard LC Win Rate
DPO	1.231 $\pm$0.045	8.43 $\pm$0.060	49.72% $\pm$0.65%	52.1% $\pm$0.65%
SimPO	1.180 $\pm$0.055	8.15 $\pm$0.072	47.80% $\pm$0.72%	51.6% $\pm$0.69%
TGDPO	1.275 $\pm$0.045	8.42 $\pm$0.060	51.40% $\pm$0.65%	52.8% $\pm$0.64%
TDPO	1.220 $\pm$0.051	8.475 $\pm$0.055	49.90% $\pm$0.71%	53.0% $\pm$0.55%
SparsePO	1.255 $\pm$0.053	8.46 $\pm$0.061	51.10% $\pm$0.63%	52.6% $\pm$0.51%
SePO	1.245 $\pm$0.055	8.40 $\pm$0.065	50.60% $\pm$0.68%	52.3% $\pm$0.65%
ADPO	**1.308** $\pm$**0.052**	**8.53** $\pm$**0.057**	**52.78%** $\pm$**0.46%**	**53.4%** $\pm$**0.59%**

Table 4: Main benchmark results for Llama-3-8B-Instruct. MT-Bench and AlpacaEval 2.0 use GPT-4-turbo as the API judge, while Arena-Hard uses GPT-4.1-mini as the API judge.

Method	FsFairX RM Mean Reward	MT-Bench Score	AlpacaEval 2.0 LC Win Rate	Arena-Hard LC Win Rate
DPO	2.15 $\pm$0.058	7.66 $\pm$0.067	50.43% $\pm$0.72%	34.9% $\pm$0.68%
SimPO	1.95 $\pm$0.063	7.60 $\pm$0.074	48.50% $\pm$0.77%	32.8% $\pm$0.73%
TGDPO	2.17 $\pm$0.052	7.75 $\pm$0.069	52.00% $\pm$0.64%	35.2% $\pm$0.62%
TDPO	2.14 $\pm$0.057	7.85 $\pm$0.061	50.00% $\pm$0.71%	35.0% $\pm$0.66%
SparsePO	2.17 $\pm$0.054	7.80 $\pm$0.068	51.40% $\pm$0.67%	35.4% $\pm$0.59%
SePO	2.13 $\pm$0.061	7.72 $\pm$0.063	50.70% $\pm$0.69%	35.1% $\pm$0.72%
ADPO	**2.20** $\pm$**0.059**	**7.89** $\pm$**0.066**	**53.06%** $\pm$**0.63%**	**36.5%** $\pm$**0.48%**

For the main evaluations, methods are trained via Full Sharded Data Parallel (FSDP) fine-tuning. Across all evaluated models, the global batch size is configured to 128 (using a per-device batch size of 4, gradient accumulation of 16, and 2 GPUs). The training runs for 1 epoch using bf16 precision, with a maximum sequence length of 2048.

We perform a hyperparameter sweep over the DPO temperature β for all methods except SimPO, evaluating $\beta \in \{0.01, 0.05, 0.1\}$. For SimPO, which uses a different reward scaling, we separately evaluate larger β values and select the best-performing setting. The hyperparameters reported below correspond to the selected configurations used for the main benchmark results in Tables 3 and 4.

- **DPO**: Serves as the reference implementation, utilizing a temperature parameter $\beta = 0.01$ and learning rate 5×10^{-7}.
- **SimPO**: Uses a reference-free length-normalized reward with $\beta = 2.5$, a target margin $\gamma = 1.375$ (calculated as $0.55 \cdot \beta$), and learning rate 1×10^{-6}.
- **TGDPO**: Employs a per-token forward-KL regularizer using $\beta = 0.01$, a token reward factor $\alpha = 0.5$, and learning rate 5×10^{-7}.
- **TDPO**: Utilizes $\beta = 0.1$, a KL weight $\alpha = 0.5$, and learning rate 5×10^{-7}.
- **SparsePO**: Uses a token-level reward and KL mask with $\beta = 0.1$, learning rate 5×10^{-7}, mask weight decay 0.01, and L1 regularization 0.001 on the mask weights.
- **SePO**: Uses the top-30% token-selection setting and a reference-free contrastive objective; the policy is trained with learning rate 5×10^{-7}.
- **ADPO**: Our proposed frequency-hard token-weighted DPO, applied with $\beta = 0.01$, learning rate 5×10^{-7}, a frequency mask threshold $\tau = 0.002$.

Table 5: Sensitivity of ADPO to the frequency threshold τ on Llama-3-8B-Instruct. Each threshold is selected to mask approximately the indicated number of vocabulary tokens. FsFairX RM reports the mean reward assigned by the FsFairX Llama-3 RM v0.1, while AlpacaEval 2.0 reports length-controlled (LC) win rate.

τ	**Masked Tokens (Approx.)**	**FsFairX RM Mean Reward**	**AlpacaEval 2.0 LC Win Rate**
0.00413	25	1.654	48.15%
0.00200	50	**2.200**	**53.06%**
0.00084	100	1.440	42.70%

Table 6: Ablation of token-selection strategies on Llama-3-8B-Instruct using approximately 50 token types. Hard-mask variants remove the indicated token set from the preference score while preserving the full context. Soft inverse-frequency weighting assigns larger weights to less frequent token types.

Method	**FsFairX RM Mean Reward**	**AlpacaEval 2.0 LC Win Rate**
DPO	2.15	50.43%
Random mask (50 tokens)	2.14	50.18%
Mid-frequency mask (50 tokens)	2.11	49.72%
Low-frequency mask (50 tokens)	2.14	49.55%
Soft inverse-frequency weights	2.16	50.62%
ADPO	**2.20**	**53.06%**

5.2 Main Results

Tables 3 and 4 compare Frequency-hard DPO (ADPO) with the considered baselines on Qwen2.5-7B-Instruct and Llama-3-8B-Instruct. On Qwen2.5-7B-Instruct, ADPO increases the FsFairX mean reward from 1.231 to 1.308 and the MT-Bench score from 8.43 to 8.53 relative to DPO. AlpacaEval 2.0 length-controlled win rate improves from 49.72% to 52.78%, while the reported Arena-Hard win rate increases from 52.1% to 53.4%. ADPO achieves the highest reported score on each metric among the compared methods.

On Llama-3-8B-Instruct, ADPO improves the FsFairX mean reward from 2.15 to 2.20 and the MT-Bench score from 7.66 to 7.89. AlpacaEval 2.0 length-controlled win rate increases from 50.43% to 53.06%, and the reported Arena-Hard win rate increases from 34.9% to 36.5%. ADPO achieves the highest reported score on each metric among the compared methods.

Table 5 examines sensitivity to the frequency threshold τ on Llama-3-8B-Instruct. Among the tested settings, $\tau = 0.002$, masking approximately 50 token types, achieves the highest mean reward and AlpacaEval 2.0 length-controlled win rate. Masking approximately 25 or 100 tokens yields lower scores on both metrics. The token-selection ablation shows that none of the alternative selection strategies provides a benefit over the standard DPO objective.

6 Conclusion

This work demonstrates that the isotropic scaling assumption in standard Direct Preference Optimization inadvertently allows a small subset of high-frequency tokens to dominate and entangle preference gradients. By introducing Anisotropic DPO (ADPO), we establish that selectively masking these frequent token types from the direct preference score—without discarding context, modifying pairs, or introducing learned parameters—effectively mitigates this interference. Extensive evaluations on Qwen2.5-7B-Instruct and Llama-3-8B-Instruct confirm that this zero-overhead frequency prior consistently improves alignment outcomes across multiple rigorous benchmarks, offering a highly efficient and transparent mechanism for robust language model preference optimization.

AI USE STATEMENT

In this work, we used generative AI tools, including ChatGPT, Claude, and Claude Code to implement methods, write and edit code, and assist in drafting the manuscript. We did not use generative AI tools to generate synthetic data, formulate , or propose our core analysis, methodology, or experimental design; these were carried out by the authors. We have reviewed all AI-assisted work: code was checked against expected behavior and the underlying data, and all AI-drafted text and result interpretations were verified against experimental outputs and revised where necessary. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International conference on artificial intelligence and statistics*, pp. 4447–4455. PMLR, 2024.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Fenia Christopoulou, Ronald Cardenas, Gerasimos Lampouras, Haitham Bou-Ammar, and Jun Wang. Sparsepo: Controlling preference alignment of llms via sparse token masks. *arXiv preprint arXiv:2410.05102*, 2024.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, et al. Ultrafeedback: Boosting language models with scaled ai feedback. *arXiv preprint arXiv:2310.01377*, 2023.

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

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization, 2024. *URL https://arxiv. org/abs/2402.01306*, 14, 2023.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Representation degeneration problem in training natural language generation models. *arXiv preprint arXiv:1907.12009*, 2019.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International conference on machine learning*, pp. 10835–10866. PMLR, 2023.

Nathan Godey, Éric Clergerie, and Benoît Sagot. Anisotropy is inherent to self-attention in transformers. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 35–48, 2024.

Aman Gupta, Shao Tang, Qingquan Song, Sirou Zhu, Jiwoo Hong, Ankan Saha, Viral Gupta, Noah Lee, Eunki Kim, Siyu Zhu, et al. Alphapo: Reward shape matters for llm alignment. *arXiv preprint arXiv:2501.03884*, 2025.

Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11170–11189, 2024.

Shawn Im, Federico Danieli, Skyler Seto, Barry-John Theobald, and Katherine Metcalf. Normalized rewards for preference optimization. *arXiv preprint arXiv:2607.16240*, 2026.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14165–14178, 2023.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

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. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. Smaug: Fixing failure modes of preference optimisation with dpo-positive. *arXiv preprint arXiv:2402.13228*, 2024.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4998–5017, 2024.

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*, 36:53728–53741, 2023.

Noam Razin, Sadhika Malladi, Adithya Bhaskar, Danqi Chen, Sanjeev Arora, and Boris Hanin. Unintentional unalignment: Likelihood displacement in direct preference optimization. In *International Conference on Learning Representations*, volume 2025, pp. 24791–24834, 2025.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *EMNLP*, 2024.

Songhao Wu, Zhongxin Chen, Yuxuan Liu, Heng Cui, Cong Li, and Rui Yan. Your unembedding matrix is secretly a feature lens for text embeddings. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 5454–5463, 2026.

Yue Wu, Zhiqing Sun, Rina Hughes, Kaixuan Ji, Yiming Yang, and Quanquan Gu. Self-play preference optimization for language model alignment. In *International Conference on Learning Representations*, volume 2025, pp. 91558–91582, 2025.

Kailai Yang, Zhiwei Liu, Qianqian Xie, Jimin Huang, Erxue Min, and Sophia Ananiadou. Selective preference optimization via token-level reward function estimation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 7043–7067, 2025.

Hee Suk Yoon, Eunseop Yoon, Mark Hasegawa-Johnson, Sungwoong Kim, and Chang D Yoo. Confpo: Exploiting policy model confidence for critical token selection in preference optimization. *arXiv preprint arXiv:2506.08712*, 2025.

Hui Yuan, Yifan Zeng, Yue Wu, Huazheng Wang, Mengdi Wang, and Liu Leqi. A common pitfall of margin-based language model alignment: Gradient entanglement. In *International Conference on Learning Representations*, volume 2025, pp. 5278–5297, 2025.

Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. Token-level direct preference optimization. *arXiv preprint arXiv:2404.11999*, 2024.

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

Mingkang Zhu, Xi Chen, Zhongdao Wang, Bei Yu, Hengshuang Zhao, and Jiaya Jia. Tgdpo: Harnessing token-level reward guidance for enhancing direct preference optimization. *arXiv preprint arXiv:2506.14574*, 2025.

A Appendix

A.1 Frequency Structure Across Tokenizers

Audit design. We evaluate the frequency prior across Mistral, Llama 3, and Qwen2.5 tokenizers using the same released Llama-3-UltraFeedback response pairs. For each tokenizer, the first 4,000 training pairs determine the fixed frequent-type set

$$\mathcal{H}_\tau = \{v : \widehat{f}(v) > \tau\}, \qquad \mathcal{K}_\tau = \mathcal{V} \setminus \mathcal{H}_\tau, \qquad \tau = 0.002. \tag{14}$$

Both responses contribute symmetrically to $\widehat{f}$. The audit applies each tokenizer's chat template, includes response EOS, and uses sequence and prompt caps of 2,048 and 1,800 tokens, respectively. These are the settings of this structural audit. The selected token IDs are frozen before evaluation on all 1,961 disjoint test pairs. Separate training pairs 4,001–8,000 calibrate the signed side associations used below. This design separates mask construction, association calibration, and held-out evaluation.

Coverage and shared occurrence mass. Let $c_i^s(v)$ count type v in response side $s \in \{w, l\}$, and let $n_{\mathcal{D}}(v) = \sum_{i \in \mathcal{D}}[c_i^w(v) + c_i^l(v)]$. For a type set $\mathcal{S}$, define its occurrence coverage and shared occurrence coverage on population $\mathcal{D}$ as

$$F_{\mathcal{D}}(\mathcal{S}) = \frac{\sum_{v \in \mathcal{S}} n_{\mathcal{D}}(v)}{\sum_{v \in \mathcal{V}} n_{\mathcal{D}}(v)}, \tag{15}$$

$$J_{\mathcal{D}}(\mathcal{S}) = \frac{\sum_{i \in \mathcal{D}} \sum_{v \in \mathcal{S}} \min\{c_i^w(v), c_i^l(v)\}}{\sum_{i \in \mathcal{D}} \sum_{v \in \mathcal{V}} \min\{c_i^w(v), c_i^l(v)\}}. \tag{16}$$

For side-normalized rates $p_{s,\mathcal{D}}(v) = \sum_{i \in \mathcal{D}} c_i^s(v) / \sum_{i \in \mathcal{D}} \sum_u c_i^s(u)$, define

$$q_{\mathcal{D}}(v) = \frac{|p_{w,\mathcal{D}}(v) - p_{l,\mathcal{D}}(v)|}{p_{w,\mathcal{D}}(v) + p_{l,\mathcal{D}}(v)}, \qquad \overline{q}_{\mathcal{D}}(\mathcal{S}) = \frac{\sum_{v \in \mathcal{S}} n_{\mathcal{D}}(v) q_{\mathcal{D}}(v)}{\sum_{v \in \mathcal{S}} n_{\mathcal{D}}(v)}. \tag{17}$$

Unobserved types contribute zero. The occurrence-weighted statistic $\overline{q}$ measures absolute token-identity side specificity.

Table 7 establishes the same structural pattern across all three tokenizers: a small frequent-type set captures a disproportionate share of chosen–rejected overlap and has substantially weaker token-identity side specificity than its complement. The masks cover 36.06–40.94% of held-out response occurrences while capturing 47.43–52.49% of their shared occurrence mass.

Reproducible preference-side associations. To measure whether side associations repeat out of sample, define the signed calibration skew

$$d_{\mathrm{cal}}(v) = \frac{p_{w,\mathrm{cal}}(v) - p_{l,\mathrm{cal}}(v)}{p_{w,\mathrm{cal}}(v) + p_{l,\mathrm{cal}}(v)}, \tag{18}$$

setting it to zero for types absent from the association-calibration population. The held-out cross-fitted score is

$$C(\mathcal{S}) = \frac{\sum_{v \in \mathcal{S}} d_{\mathrm{cal}}(v)[p_{w,\mathrm{test}}(v) - p_{l,\mathrm{test}}(v)]}{\sum_{v \in \mathcal{S}}[p_{w,\mathrm{test}}(v) + p_{l,\mathrm{test}}(v)]}. \tag{19}$$

Positive values indicate that the direction of a token's side association reproduces on disjoint pairs. We resample evaluation preference pairs jointly across the frequent and retained groups for 10,000 bootstrap draws, keeping the training-derived mask and calibration associations fixed. The resulting intervals quantify evaluation-pair uncertainty.

Retained types have a larger cross-fitted side signal under every tokenizer, and each paired interval for the difference is above zero (Table 8). Together, the coverage and cross-fitted analyses support a tokenizer-replicated frequency prior: frequent types concentrate shared occurrences, whereas retained types carry stronger reproducible token-identity preference-side associations per unit group mass.

Table 7: Held-out coverage and side specificity of training-derived frequency masks. Coverage is computed from response occurrences; shared coverage uses within-pair multiset overlap. The same test pairs are tokenized separately for each row.

Tokenizer	$\lvert\mathcal{H}_\tau\rvert$	$F(\mathcal{H}_\tau)$ (%)	$J(\mathcal{H}_\tau)$ (%)	$\overline{q}(\mathcal{H}_\tau)$	$\overline{q}(\mathcal{K}_\tau)$
Mistral	53	40.94	52.49	0.0126	0.1033
Llama 3	50	36.06	47.43	0.0133	0.1406
Qwen2.5	52	37.77	49.22	0.0135	0.1404

Table 8: Cross-fitted preference-side signal. All entries are scaled by 10^3. Intervals are paired percentile bootstrap intervals for the retained-minus-frequent difference, computed separately per tokenizer.

Tokenizer	$C(\mathcal{H}_\tau)$	$C(\mathcal{K}_\tau)$	Difference	95% CI
Mistral	0.326	1.234	0.908	$[0.382, 1.439]$
Llama 3	0.315	1.602	1.287	$[0.563, 2.022]$
Qwen2.5	0.292	1.558	1.266	$[0.546, 1.981]$

Exact masked-token inventories. Table 9 gives every selected token ID and its tokenizer-piece string for the three training-derived masks. Strings are JSON-escaped, with non-ASCII characters represented by Unicode escapes. For example, `\u2581` is the Mistral word-boundary marker; `\u0120`, `\u010a`, and `\u0109` denote the byte-level space, newline, and tab markers in the other tokenizers. These are vocabulary pieces, not separately decoded text fragments. Each tokenizer's columns follow descending calibration frequency; positions across tokenizers do not imply corresponding token identities.

Table 9: Complete frequent-type inventories for the structural audit. Token IDs and escaped vocabulary pieces preserve whitespace, byte, and special-token distinctions.

Mistral		Llama 3		Qwen2.5	
ID	Vocabulary piece	ID	Vocabulary piece	ID	Vocabulary piece
13	"<0x0A>"	11	","	11	","
28723	"."	279	"\u0120the"	279	"\u0120the"
28725	","	323	"\u0120and"	323	"\u0120and"
272	"\u2581the"	13	"."	13	"."
304	"\u2581and"	311	"\u0120to"	311	"\u0120to"
298	"\u2581to"	264	"\u0120a"	264	"\u0120a"
264	"\u2581a"	315	"\u0120of"	315	"\u0120of"
28705	"\u2581"	220	"\u0120"	220	"\u0120"
302	"\u2581of"	25	":"	16	"1"
28747	":"	304	"\u0120in"	15	"0"
28740	"1"	9	"*"	25	":"
28742	"’"	627	".\u010a"	17	"2"
28733	"-"	374	"\u0120is"	304	"\u0120in"
28734	"0"	198	"\u010a"	9	"*"
28713	"s"	3146	"\u0120**"	624	".\u010a"
297	"\u2581in"	320	"\u0120("	374	"\u0120is"
28750	"2"	382	".\u010a\u010a"	198	"\u010a"
28736	"*"	430	"\u0120that"	3070	"\u0120**"
349	"\u2581is"	369	"\u0120for"	320	"\u0120("
325	"\u2581("	596	"’s"	382	".\u010a\u010a"
619	"\u2581**"	449	"\u0120with"	429	"\u0120that"
348	"**"	330	"\u0120\""	369	"\u0120for"
354	"\u2581for"	477	"\u0120or"	594	"’s"
369	"\u2581that"	96618	"**:"	448	"\u0120with"
28731	")"	16	"1"	330	"\u0120\""
395	"\u2581with"	334	"**"	18	"3"
9189	"**:"	649	"\u0120can"	476	"\u0120or"
345	"\u2581\""	17	"2"	95518	"**:"
28770	"3"	389	"\u0120on"	334	"**"
442	"\u2581or"	439	"\u0120as"	646	"\u0120can"
28730	"_"	499	"\u0120you"	389	"\u0120on"
28832	"`"	578	"\u0120The"	438	"\u0120as"
541	"\u2581can"	284	"\u0120="	20	"5"
356	"\u2581on"	197	"\u0109"	498	"\u0120you"
390	"\u2581as"	701	"\u0120your"	576	"\u0120The"

Continued on the next page

Table 9 continued

Mistral ID	Mistral Vocabulary piece	Llama 3 ID	Llama 3 Vocabulary piece	Qwen2.5 ID	Qwen2.5 Vocabulary piece
28782	"5"	18	"3"	19	"4"
415	"\u2581The"	262	"\u0120\u0120\u0120"	284	"\u0120="
368	"\u2581you"	1595	"\u0120`"	197	"\u0109"
12	"<0x09>"	1473	":\u010a\u010a"	697	"\u0120your"
28781	"4"	527	"\u0120are"	262	"\u0120\u0120\u0120"
327	"\u2581="	433	"\u0120it"	1565	"\u0120`"
1552	"\u2581`"	8	")"	1447	":\u010a\u010a"
574	"\u2581your"	57277	"**\u010a\u010a"	525	"\u0120are"
288	"ing"	387	"\u0120be"	432	"\u0120it"
227	"<0xE0>"	128001	"<|end_of_text|>"	8	")"
28732	"("	1	"\""	56177	"**\u010a\u010a"
2287	"\u2581\u2581\u2581"	63	"`"	387	"\u0120be"
28739	"\""	358	"\u0120I"	151645	"<|im_end|>"
460	"\u2581are"	555	"\u0120by"	1	"\""
378	"\u2581it"	19	"4"	63	"`"
347	"\u2581be"			358	"\u0120I"
315	"\u2581I"			553	"\u0120by"
2	"</s>"				

A.2 Controlled Token Selection

An occurrence-budget control. Vocabulary cardinality and response-occurrence coverage describe different properties of a selector. A fixed number of randomly selected vocabulary types need not cover the same number of response positions as the most frequent types. We therefore complement type-count comparisons with a structural control that matches selected occurrence mass.

The Qwen2.5 control uses the first 4,000 training pairs of the same pinned Llama-3-UltraFeedback release, the Qwen2.5-1.5B-Instruct tokenizer (revision `989aa798`), chat formatting, response EOS, and the 2,048/1,800 sequence/prompt caps. At $\tau = 0.002$, the frequency selector contains 52 types and covers 1,194,145 of 3,176,166 response occurrences (37.60%). We draw 1,000 randomized selections from observed vocabulary types without replacement, matching this occurrence budget rather than the number of types. Specifically, each draw permutes the observed types and accepts a type only when its complete occurrence count fits the remaining budget; types that would exceed it are skipped. Selection stops when the remaining budget reaches zero. Thus the controls are binary type-level selections with exactly matched occurrence counts. Their vocabulary sizes are allowed to vary. The randomization seed is 42.

We compare shared coverage J, occurrence-weighted side specificity $\overline{q}$, and the shared enrichment

$$E(\mathcal{S}) = \frac{J(\mathcal{S})}{F(\mathcal{S})}. \tag{20}$$

An enrichment above one means that the selected occurrence mass contains a disproportionately large share of chosen–rejected overlap. The randomized selections define an empirical reference distribution for each statistic. One-sided Monte Carlo values use the upper tail for J and E and the lower tail for $\overline{q}$, with the plus-one correction $(1+b)/(1+1{,}000)$, where b counts randomized statistics at least as extreme as the frequency selector in the tested direction. At the matched occurrence budget, frequency selection captures more shared chosen–rejected mass and substantially weaker token-identity side specificity than randomized selection (Table 10). Its shared coverage is 49.03%,

Table 10: Frequency selection and occurrence-matched randomized selection on the training-calibration population. The range is the central empirical range across randomized selections, not a confidence interval for a trained-model effect. One-sided Monte Carlo values are unadjusted.

Statistic	Frequency	Random mean	Random 95% range	p
Selected types	52	20,911.4	[16,756.4, 24,749.7]	—
Occurrence coverage (%)	37.60	37.60	[37.60, 37.60]	—
Shared coverage (%)	49.03	37.46	[34.62, 40.71]	0.001
Shared enrichment	1.304	0.996	[0.921, 1.083]	0.001
Side specificity	0.0152	0.0748	[0.0625, 0.0864]	0.001

compared with a random mean of 37.46%, and its shared enrichment is 1.304. Thus the observed structural concentration is not explained by the amount of selected occurrence mass alone.

The held-out replication in Appendix A.1 and this training-calibration randomization address complementary questions: the former establishes that the structural pattern reproduces on disjoint responses across tokenizers, and the latter tests the specificity of frequency selection at a fixed occurrence budget. Both analyses use token counts rather than model optimization. They support frequency as a transparent selector of shared, weakly side-specific response content for the score-masking objective.