The Reference-Conditioning Trap: A Population Bound on Mis-Ranking Corrections under Fixed-Reference DPO

Liz Lemma Future Detective

January 22, 2026

Abstract

Modern preference learning pipelines (DPO/RLHF) are explicitly regularized toward a reference model π_{ref}. Chen et al. (NeurIPS 2024) show that this reference conditioning makes correcting even mild ranking errors difficult: for a datapoint with reference log-ratio c, DPO flips the ranking if and only if the DPO loss falls below $-\log\sigma(\beta c)$. We push this pointwise characterization into an economics-style population limit that is directly actionable for 2026-era alignment governance. We introduce a clean finite-optimization/early-stopping floor $\underline{L}(B)$ parameterized by compute budget B, and derive a closed-form upper bound on the fraction of mis-ranked preference pairs that any fixed-reference DPO procedure can correct before overfitting/degeneration. The bound depends only on β, $\underline{L}(B)$, and the upper tail of the reference log-ratio distribution c. This yields a testable "alignment-cap frontier": when π_{ref} assigns too much likelihood to dispreferred completions (large positive c), no feasible reduction in loss at early stopping can flip those pairs. We outline empirical validation by estimating c-distributions and per-example losses across checkpoints, then predicting realized ranking flips across models/datasets. The results provide an interpretable diagnostic for when alignment investment must shift from 'more DPO' to improving the reference, refreshing the dataset on-policy, or redesigning objectives.

Table of Contents

3. 3. Setup and notation: aggregated preference data; reference log-ratio distribution; defining mis-ranked set $\mathcal{M}$; per-point DPO loss and the flip condition.

4. 4. A tractable finite-optimization model: early-stopping as a constraint; defining an empirical or theoretical floor $\underline{L}(B)$; discussion of how $\underline{L}(B)$ can be estimated from training traces.

5. 5. Main result (population bound): derivation of $\mathrm{Corr}(B, \beta) \leq F_{\mathcal{M}}(c^*(B, \beta))$; interpretation as an 'uncorrectable tail mass' statement; sensitivity to β and reference quality.

6. 6. Extensions (optional but clean): (i) heterogeneous learnability / knapsack allocation and when closed forms fail; (ii) bounds under average-loss constraints via Markov-type arguments; (iii) variants such as γ-scaled reference terms.

7. 7. Empirical validation plan: estimating c, L_i, flips; predicting correction ceilings across datasets; comparing to observed flip rates under early stopping; robustness checks (length normalization, ties, alternative ranking definitions).

8. 8. Implications for 2026 alignment governance: when 'more preference tuning' is futile; when to invest in better references, on-policy refresh, or protocol changes; interpret β as inertia/adjustment cost; concrete diagnostics for practitioners.

9. 9. Discussion and limitations: relationship to win rate divergence; what the bound does not capture; how to relax the floor assumption; future directions (dynamic references, market-like auditing).

1 Introduction

A recurring empirical pattern in modern preference learning is that we can often improve *average* user-facing quality while still failing to repair the most salient "alignment gaps"—the specific prompts where a deployed model reliably chooses a completion that users (or internal raters) judge worse. Practitioners experience this as a kind of brittleness: training appears to move many logits in the right direction, yet some misbehaviors remain stubbornly fixed, even when they are clearly labeled in the preference data. In this paper we formalize one mechanism behind this phenomenon for fixed-reference Direct Preference Optimization (DPO): correcting a mis-ranked pair can be viewed as paying an *adjustment cost* that grows with how strongly the reference model prefers the wrong answer, and finite compute budgets impose an effective floor on how much adjustment cost we can pay.

Our motivating lens is economic. When an agent is anchored to a default behavior (here, the reference policy), moving away from that default is not free: it requires optimization effort, and at finite horizons we should expect "sticky" outcomes where some mistakes persist. This perspective is especially natural in RLHF-style pipelines: we begin from a pretrained model that already encodes a large amount of capability and stylistic prior, then we apply preference learning as a comparatively small adjustment step that must not destroy the underlying competence. In that regime, the training algorithm is implicitly solving a constrained problem: it must improve preference satisfaction without drifting too far, and it must do so under a limited optimization budget. The central question is therefore not merely whether DPO is consistent in the limit of infinite compute, but what it can *reliably* change at the checkpoints we actually deploy.

Fixed-reference preference learning makes the anchoring particularly explicit. DPO, as commonly implemented, optimizes a policy relative to a *frozen* reference model. The reference enters the objective through a log-ratio term that reweights how hard it is to make the policy prefer a labeled winner over a labeled loser. Intuitively, if the reference already prefers the winner, DPO only needs to "nudge" the policy in the same direction; if the reference prefers the loser, DPO must fight uphill against the reference odds. This uphill case is exactly what we care about from an alignment standpoint: these are datapoints where the deployed base model is biased toward the wrong completion even though the aggregated preference label indicates the opposite. These mis-ranked pairs are the ones that, in practice, constitute the most visible safety and product risks.

The complication is that we never optimize the DPO objective to zero. Real training uses early stopping, regularization, and finite compute; moreover, we often select checkpoints by validation heuristics that prioritize broad generalization rather than aggressively fitting the hardest (and often rare) mis-ranked points. This means that, even on the training dataset, there is

a residual per-example loss that does not go away. We model this residual effect via a finite-optimization "loss floor" on the mis-ranked subset: after running a training-and-stopping procedure with budget B, the attained checkpoint cannot push every mis-ranked point below some minimum achievable loss level. While this floor can be motivated by optimization limits, it also captures a practical governance reality: organizations typically have explicit caps on training time, and they adopt conservative stopping rules to reduce overfitting, mode collapse, or capability degradation. In short, there is an operationally meaningful sense in which "we stop before we have fixed everything."

Our main message is that the interaction between (i) the reference model's mis-ranking strength and (ii) a finite loss floor yields a sharp ceiling on the fraction of mis-ranked points that can be corrected by fixed-reference DPO. Concretely, the reference log-ratio for a labeled pair induces a *difficulty threshold*: to flip the ranking at a checkpoint, the per-point DPO loss must be driven below a value that decreases as the reference becomes more confident in the wrong completion. Therefore, if early stopping implies a lower bound on the attainable loss on mis-ranked points, only those mis-ranked points whose thresholds lie above that floor are even *feasible* to correct. Aggregating across datapoints turns this into a population statement: the corrected fraction is upper bounded by the cumulative mass of "mild" mis-rankings, i.e. those where the reference's bias in favor of the loser is not too large.

This framing clarifies why alignment gaps can persist even when we have labeled data directly addressing them. A preference dataset can contain many mis-ranked pairs, but if a nontrivial portion of those pairs sit in the heavy upper tail of the reference log-ratio distribution, then they are effectively "locked in" by the combination of anchoring and finite optimization. Under fixed-reference training, we should not expect these points to flip unless we either (a) spend substantially more compute (so the loss floor decreases), (b) reduce the strength of the anchoring (e.g. via hyperparameters), or (c) change the reference itself over time (e.g. iterative methods). Importantly, these options correspond to real tradeoffs faced by deployers: more compute is costly; weaker anchoring can increase distributional shift and degrade capabilities; and changing the reference can complicate evaluation, reproducibility, and auditing.

Our contributions are threefold. First, we connect the pointwise difficulty phenomenon in DPO to a simple, testable *alignment-cap frontier* under finite compute: a closed-form correction ceiling determined by the reference log-ratio distribution and the optimization floor. Second, we provide comparative statics that match practitioner intuitions but make them precise: increasing the DPO inverse-temperature parameter can tighten the feasibility thresholds and thus reduce the maximum correctable mass at a fixed floor; increasing compute can expand the feasible region by lowering the floor; and

improving the reference model reshapes the mis-ranking tail, increasing the achievable correction rate even at the same compute. Third, we highlight safety and governance implications: if a training method has an intrinsic ceiling on correcting the most severe mis-rankings under realistic budgets, then evaluation protocols should explicitly measure the residual uncorrectable tail and avoid over-interpreting aggregate win-rate gains as evidence of robust correction on worst-case prompts.

The rest of the paper develops this argument in a step-by-step way. In the next section we recall the DPO objective and the distinction between on-policy win rate and offline ranking correctness, and we summarize the key threshold phenomenon that makes mis-ranked points qualitatively harder than correctly-ranked points. We then formalize the finite-optimization floor assumption and derive the population correction bound, interpret it as a frontier between compute, anchoring, and correctable mis-rankings, and discuss extensions that allow heterogeneous learnability across datapoints. Finally, we outline an empirical protocol for checking whether observed training runs respect the predicted ceiling—and what it would mean, from both a scientific and a governance perspective, to observe systematic violations.

2 Background and motivation

2.1 From RLHF to fixed-reference DPO

Most deployed preference-learning stacks can be viewed as variations on a common theme: we wish to update a pretrained language model to better match human judgments while retaining the broad competence encoded by pretraining. In RLHF, this is often expressed as maximizing an (implicit) reward model subject to a Kullback–Leibler (KL) penalty to a *reference* policy. The KL term is not merely a technical convenience; it is the mechanism by which organizations operationalize a safety–capability tradeoff. It discourages large distributional shifts, reduces the risk of reward hacking, and provides a knob for controlling how aggressively we overwrite the pretrained prior.

Direct Preference Optimization (DPO) can be understood as a particularly clean instantiation of this anchored update, where the optimization problem is written directly in terms of pairwise preferences and a fixed reference model. Concretely, DPO posits (or is equivalent to) a Bradley–Terry style likelihood for preferences under a policy π_θ, and then expresses the learned policy as a multiplicative reweighting of the reference policy by an exponentiated reward. In the resulting objective, the reference is frozen and appears inside the per-example likelihood as an additive log-odds correction. This "reference log-odds" term is what makes DPO attractive from an engineering standpoint: we obtain a stable, supervised-learning-style loss, but we retain the anchoring behavior of KL-regularized RLHF.

The key subtlety, and the one that motivates our analysis, is that anchoring is asymmetric across datapoints. When the reference already leans toward the preferred completion, the update is "downhill" and the optimization only needs to reinforce an existing preference. When the reference leans toward the dispreferred completion, the update is "uphill": the optimizer must fight against the reference odds, and the amount of work required depends on *how strongly* the reference is miscalibrated. This asymmetry is easy to miss if we only look at average training loss curves, but it becomes central once we care about stubborn, safety-relevant errors that persist across fine-tuning runs.

2.2 Two notions of "improving preferences": win rate vs. ranking correctness

A second motivation for our setup is that "preference satisfaction" is measured in multiple, non-equivalent ways in practice. A common online or quasi-online metric is *win rate*: we sample responses from the trained policy and from a baseline (often the reference or a previous checkpoint), then ask raters which response is better. Win rate is the metric that product teams often care about because it captures user-visible improvements under the policy's own sampling distribution.

Our theoretical results, by contrast, focus on an *offline* notion: whether the learned policy assigns higher conditional probability to the labeled winner than to the labeled loser on the same prompt. We refer to this as ranking correctness on the preference dataset. This metric is not a substitute for win rate, but it is an important diagnostic for two reasons. First, it isolates the specific mechanism we study: the way fixed-reference objectives encode an *odds constraint* on how hard it is to reverse a mis-ranking. Second, it is the natural quantity that connects to per-example DPO loss: on a fixed pair (x, y^w, y^ℓ), the model "fixes" the error precisely when it assigns at least as much probability mass to y^w as to y^ℓ. In other words, ranking correctness makes explicit whether training has actually reversed the reference model's preference on that exact disagreement pair, rather than merely producing alternative samples that happen to win more often.

This distinction matters operationally. A training run can increase win rate by improving fluency, harmlessness style, or generic helpfulness while still failing to reverse the ranking on the most egregious disagreement pairs. From a safety perspective, these disagreement pairs are often the ones we care about most: they are where the base model is *systematically* attracted to a completion that human labelers prefer less (e.g., a subtly unsafe instruction-following behavior, or a misleading answer that sounds confident). If our evaluation collapses these cases into an average win rate, we risk mistaking broad style improvements for robust correction of the problematic tail.

2.3 Chen et al.: why mis-ranked pairs are intrinsically "hard" under DPO

Chen et al. make the above asymmetry precise and document it empirically. Their core observation is that, for DPO with a fixed reference, there exists a pointwise "difficulty threshold" for flipping the ranking on any given preference pair, and this threshold depends directly on the reference model's relative preference for the two completions. Intuitively, the reference log-odds act like an offset in the logistic loss: if the reference assigns higher probability to the loser than to the winner, then the optimizer must produce a *larger* change in the policy's own log-odds before the logistic term becomes confident in the correct direction.

At a high level, the per-pair DPO contribution takes the form of a negative log-likelihood,

$$-\log \sigma\Big(\beta\Big(\log \pi_\theta(y^w \mid x) - \log \pi_\theta(y^\ell \mid x) + (\text{reference offset})\Big)\Big),$$

where $\beta > 0$ is the inverse-temperature hyperparameter controlling how sharply preferences are enforced. Theorem 4.1 in Chen et al. shows that the event "the policy ranks y^w above y^ℓ" is equivalent to the per-example loss being below a threshold that is a deterministic function of the reference offset. The important qualitative property is monotonicity: as the reference becomes *more* confident in the wrong completion, the threshold becomes *smaller*, meaning the loss must be driven closer to zero to flip the ranking.

This immediately suggests an "alignment gap" mechanism that does not rely on label noise or representational limits. Even if the preference dataset is correct and internally consistent, and even if the model class is expressive enough, the optimization dynamics can still fail to flip many mis-ranked pairs because doing so requires pushing their individual losses below extremely stringent thresholds. Chen et al. report that, in realistic regimes, the resulting ranking accuracy on the dataset can remain surprisingly low, especially on the subset of pairs where the reference disagrees with the preference label. In our terms, the model improves broadly but leaves a residue of uncorrected mis-rankings.

2.4 Why this becomes a compute-and-governance issue

Theorem 4.1 is a pointwise statement; on its own it does not yet say what fraction of mis-rankings will remain in a full training run. The missing ingredient is that real training is compute-limited and deliberately stopped early. Early stopping is not merely about speed; it is often a governance and safety control. Teams stop when validation metrics plateau, when divergence from the reference becomes concerning, or when downstream capability regressions appear. Consequently, there is an empirical sense in which training can only push per-example losses "so far down" before we halt.

Once we combine (i) a per-example threshold that can be arbitrarily stringent for strongly mis-ranked pairs with (ii) a finite, operationally imposed limit on how low losses can go at the chosen checkpoint, we obtain a sharp prediction: only the "mild" reference disagreements are even feasible to correct under fixed-reference DPO at the compute budgets we actually use. The hard tail—pairs for which the reference heavily favors the loser—becomes effectively locked in.

This is the motivating bridge to our main result. In the next section we formalize the preference dataset, the fixed reference offsets, and the mis-ranked subset, and then we translate Chen et al.'s pointwise threshold into a population-level bound: an explicit ceiling on the fraction of reference mis-rankings that can be corrected at an early-stopped checkpoint, expressed in terms of the reference log-odds distribution and an optimization-induced loss floor.

3 Setup and notation: aggregated preferences, reference log-odds, and the flip condition

We work in the standard offline preference-learning regime: a fixed, aggregated dataset of pairwise comparisons, a fixed *reference* model that supplies the anchoring distribution, and a trainable policy that is updated by minimizing a supervised loss. The goal of this section is to make explicit the object that, in our view, governs most of the "hard cases" under fixed-reference DPO: the *distribution* of reference log-odds on the subset of pairs where the reference disagrees with the majority label.

Aggregated preference data. Let

$$D \;=\; \{(x_i, y_i^w, y_i^\ell)\}_{i=1}^n$$

denote an aggregated preference dataset. Here x_i is the prompt (or conversational context) and (y_i^w, y_i^ℓ) is an ordered pair of completions, where y_i^w is the *winner* and y_i^ℓ is the *loser* under the dataset's aggregation rule (e.g., majority vote over raters, possibly after rater-quality weighting). We treat this ordering as fixed: whatever uncertainty or heterogeneity exists at the individual-rater level has already been compressed into the deterministic label "y^w is preferred to y^ℓ." This is the regime in which DPO is most commonly deployed: the training loop does not revisit the underlying preference elicitation, and the optimization problem is defined entirely by the aggregated pairs.

We write $\pi_{\mathrm{ref}}(\cdot \mid x)$ for the reference policy and $\pi_\theta(\cdot \mid x)$ for the trainable policy. Both are conditional language models over completions given context x. Throughout, we fix an inverse-temperature hyperparameter $\beta > 0$, which controls how sharply the loss penalizes disagreements with the preference

labels (equivalently, how strongly preference odds are enforced relative to the anchoring effect of the reference).

Reference log-ratio as a per-example "headwind." For each datapoint i, define the reference log-ratio

$$c_i \;\equiv\; \log \frac{\pi_{\mathrm{ref}}(y_i^\ell \mid x_i)}{\pi_{\mathrm{ref}}(y_i^w \mid x_i)} \;=\; \log \pi_{\mathrm{ref}}(y_i^\ell \mid x_i) - \log \pi_{\mathrm{ref}}(y_i^w \mid x_i).$$

This scalar is an easily-computable summary of how the fixed reference views the labeled pair. It has a direct operational interpretation. If $c_i < 0$, the reference already assigns higher probability to the dataset winner y_i^w than to the loser y_i^ℓ; preference training is (locally) "with the grain" of the reference. If $c_i > 0$, the reference prefers the loser; preference training must overcome a headwind whose magnitude is exactly c_i in log-odds units. In practice, this headwind aggregates multiple sources of difficulty: the reference may be systematically miscalibrated on certain topics; the pair may be stylistically atypical relative to pretraining; or the loser may exploit a strong prior (e.g., confident tone) that the reference overweights.

Because our interest is not just in whether such points exist but in *how many* exist at each difficulty level, we will repeatedly consider the empirical distribution of $\{c_i\}$, and especially the distribution conditional on $c_i > 0$. This emphasis is deliberate: the *tail* of large positive c_i corresponds to pairs where the reference is confidently wrong, and these are precisely the pairs that tend to be most stubborn under anchored objectives.

The mis-ranked subset. Define the set of reference-mis-ranked datapoints

$$\mathcal{M} \;\equiv\; \{i \in \{1, \ldots, n\} : c_i > 0\}.$$

On $\mathcal{M}$, the preference label and the reference ranking disagree: $\pi_{\mathrm{ref}}(y_i^\ell \mid x_i) > \pi_{\mathrm{ref}}(y_i^w \mid x_i)$ even though the aggregated human judgment says $y_i^w \succ y_i^\ell$. While many training curves and ablations report aggregate statistics over all of D, our results will be "conditioned" on $\mathcal{M}$, because this is where DPO must reverse a preference rather than merely amplify one. When $|\mathcal{M}|$ is non-negligible, the behavior on this subset is often what determines whether preference training corrects systematic failure modes or merely polishes already-good behavior.

We will write $F_{\mathcal{M}}$ for the cumulative distribution function (CDF) of c restricted to $\mathcal{M}$. In finite samples, the natural estimator is the empirical CDF

$$\widehat{F}_{\mathcal{M}}(t) \;=\; \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \mathbf{1}[c_i \le t],$$

and in large-sample discussions we treat $F_{\mathcal{M}}(t) \approx \mathbb{P}[c \le t \mid c > 0]$. Intuitively, $F_{\mathcal{M}}$ answers: among the reference disagreements, what fraction are "mild" (small c) versus "severe" (large c)?

Per-example DPO loss. For each datapoint i, define the DPO per-example loss

$$L_i(\theta) \;\equiv\; -\log \sigma\left(\beta\left(\log \frac{\pi_\theta(y_i^w \mid x_i)}{\pi_\theta(y_i^\ell \mid x_i)} + c_i\right)\right),$$

where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function. It is helpful to separate the θ-dependent term from the fixed offset. Let

$$\Delta_i(\theta) \;\equiv\; \log \frac{\pi_\theta(y_i^w \mid x_i)}{\pi_\theta(y_i^\ell \mid x_i)}.$$

Then $L_i(\theta) = -\log \sigma(\beta(\Delta_i(\theta) + c_i))$. The role of the reference is entirely captured by the additive constant c_i. In particular, for fixed c_i, the loss is a strictly decreasing function of $\Delta_i(\theta)$: increasing the policy's log-odds in favor of the winner monotonically decreases the loss. However, the *baseline* difficulty differs sharply across i because the point $\Delta_i(\theta) = 0$ (where the policy is indifferent between y_i^w and y_i^ℓ) corresponds to different loss values depending on c_i.

The empirical DPO objective is typically the average loss $\widehat{L}_{\mathrm{DPO}}(\theta; \beta) = \frac{1}{n}\sum_{i=1}^n L_i(\theta)$ (possibly with additional regularization), but our subsequent arguments do not require committing to a particular optimizer, schedule, or batching strategy; we only use the fact that DPO training attempts to reduce these per-example losses subject to the practical constraints of finite compute and early stopping.

Ranking correctness and the flip threshold. The offline notion of "fixing" a reference mis-ranking is that the trained policy assigns at least as much probability to the labeled winner as to the labeled loser. Formally, define the ranking correctness indicator

$$R_i(\theta) \;\equiv\; \mathbf{1}\Big[\pi_\theta(y_i^w \mid x_i) \ge \pi_\theta(y_i^\ell \mid x_i)\Big] \;=\; \mathbf{1}[\Delta_i(\theta) \ge 0].$$

Chen et al. (Theorem 4.1) show that this event is equivalent to a simple per-example loss threshold. To see the mechanism, evaluate the loss at the knife-edge $\Delta_i(\theta) = 0$:

$$\tau_i \;\equiv\; -\log \sigma(\beta c_i).$$

Because $L_i(\theta)$ is strictly decreasing in $\Delta_i(\theta)$, we have the equivalence

$$R_i(\theta) = 1 \quad\Longleftrightarrow\quad \Delta_i(\theta) \ge 0 \quad\Longleftrightarrow\quad L_i(\theta) \le \tau_i.$$

Thus each pair comes with a deterministic "flip threshold" τ_i, and the threshold is itself a monotone function of the reference headwind c_i. In particular, on $\mathcal{M}$ where $c_i > 0$, we have $\sigma(\beta c_i) > \frac{1}{2}$ and hence $\tau_i \in (0, \log 2)$. Moreover, as c_i increases, $\tau_i = -\log \sigma(\beta c_i)$ decreases toward 0. The more confidently the reference prefers the loser, the closer to *zero* the per-example loss must be pushed before the policy even becomes indifferent between the two completions, let alone decisively prefers the winner.

This is the key link we will exploit: the distribution of c_i on $\mathcal{M}$ induces a corresponding distribution of thresholds τ_i, and any practical constraint on how far optimization can reduce per-example losses immediately becomes a constraint on how many of these thresholds can be satisfied. The next section formalizes this constraint via an early-stopping (or finite-optimization) loss floor and uses it to convert the pointwise threshold into a population-level ceiling on the fraction of mis-rankings that can be corrected.

A finite-optimization viewpoint. In deployments, we rarely run preference optimization to full convergence on the offline dataset. Compute is bounded (by wall-clock, tokens, or optimizer steps), and we typically impose an explicit stopping rule to manage overfitting, distribution shift, or unacceptable divergence from the reference. This motivates treating the training loop not as an unconstrained minimization of the empirical DPO objective, but as a *budgeted* procedure: given a compute budget B, an algorithm produces a checkpoint θ_B that is merely *reachable* under that budget.

Formally, we let $\mathcal{A}(B)$ denote a training-and-stopping procedure (optimizer, schedule, batching, and stopping rule) that returns a checkpoint θ_B. One can think of $\mathcal{A}(B)$ as inducing a (possibly random) path $\{\theta_t\}_{t \le T(B)}$ with $T(B)$ the maximum number of updates permitted by budget, and then selecting a stopping time $t_B \le T(B)$ (e.g., the first time validation loss stops improving, or the iterate with minimum held-out loss). The key modeling move is to summarize all these operational details by a single scalar that captures what finite optimization can (and cannot) accomplish on the *hard* subset $\mathcal{M}$.

Early stopping as a per-example loss constraint. The flip condition from the previous section is pointwise: for each $i \in \mathcal{M}$, correcting the reference mis-ranking requires driving $L_i(\theta)$ below its threshold $\tau_i = -\log \sigma(\beta c_i)$. The obstacle is that, under finite compute and practical stopping rules, per-example losses do not decrease arbitrarily. In particular, the mis-ranked examples $\mathcal{M}$ often exhibit slow improvement because they require the model to *undo* a strong prior inherited from π_{ref} (large positive c_i), and because gradients from different examples interfere in shared parameters.

We encode this operational limitation via a *loss floor* on $\mathcal{M}$. We assume there exists a (typically decreasing) function $\underline{L}(B) \ge 0$ such that the selected

checkpoint $\theta_B = \mathcal{A}(B)$ satisfies

$$\forall i \in \mathcal{M} : \qquad L_i(\theta_B) \;\geq\; \underline{L}(B).$$

Intuitively, $\underline{L}(B)$ is the smallest per-example DPO loss level that finite training reliably reaches *on the mis-ranked points* before the stopping rule halts optimization. This is a deliberately coarse abstraction: it forgets which examples are hardest, and it ignores heterogeneity in optimization dynamics. Its role is to give us a tractable, pessimistic handle on what early stopping implies for the feasibility of flipping many mis-rankings.

It is important to stress what this floor is (and is not). It is *not* a statement that the average loss cannot be reduced further, nor that some examples cannot be fit. Rather, it is a statement that under the deployed training-and-stopping procedure, there remains a nontrivial subset of mis-ranked examples whose losses are not pushed below a certain level, and we summarize that phenomenon by a single bound applied uniformly on $\mathcal{M}$. When the uniformity is too strong, one can relax it (e.g., to quantile floors); we return to this in our empirical discussion.

Why should a floor exist? From a mechanistic perspective, several distinct constraints can generate an effective $\underline{L}(B)$ even when the model class is expressive. First, *optimization time*: if we stop after T gradient steps, there is a hard limit on the maximum change in log-odds $\Delta_i(\theta)$ achievable for the slowest-moving datapoints, especially under small learning rates or conservative schedules. Second, *implicit or explicit regularization*: many DPO implementations include KL penalties, weight decay, reference-mix baselines, or other stabilizers that prevent large departures from π_{ref}; these stabilizers are often tuned precisely to avoid catastrophic degradation, but they also constrain the attainable per-example loss on examples with large headwinds c_i. Third, *gradient conflict and shared capacity*: the parameters that would reduce L_i for a particular mis-ranked pair may simultaneously increase losses elsewhere, so a global optimizer that reduces the average loss may leave some individual losses relatively high at the selected checkpoint. Finally, *stopping rules are welfare-driven, not feasibility-driven*: practitioners stop when generalization, safety metrics, or divergence constraints look acceptable. That stopping time need not coincide with the time at which hard mis-ranked examples cross their flip thresholds.

For our purposes, the detailed cause is less important than the observable fact: in many runs, the left tail of $\{L_i(\theta_B) : i \in \mathcal{M}\}$ does not approach 0, and the worst-case (or near-worst-case) losses on $\mathcal{M}$ remain bounded away from 0 at early stopping.

Interpreting $\underline{L}(B)$ as a compute–capability frontier. We view $\underline{L}(B)$ as a reduced-form description of a compute–capability tradeoff on the mis-ranked subset. As B increases (more steps, larger batches, more tokens,

longer schedules), $\underline{L}(B)$ should weakly decrease, reflecting improved optimization. However, in realistic regimes it may exhibit diminishing returns: after a point, additional compute primarily reduces already-small losses or improves easy examples, while the hardest examples on $\mathcal{M}$ remain stuck due to regularization or representation limits. This is exactly the regime where a tail-based impossibility statement becomes informative: the problem is not that we cannot fit *anything*, but that we cannot fit *the high-headwind tail* without paying additional costs (compute, divergence, or degradation elsewhere).

Estimating the floor from training traces. The floor can be treated either as a theoretical constraint (for a worst-case guarantee) or as an empirically estimable quantity (for falsifiable prediction). Empirically, we can log per-example losses during training and compute $L_i(\theta_B)$ for each $i \in \mathcal{M}$ at the selected checkpoint. A conservative estimator that literally satisfies the uniform constraint is

$$\widehat{\underline{L}}_{\min}(B) \;\equiv\; \min_{i \in \mathcal{M}} L_i(\theta_B),$$

since by construction $L_i(\theta_B) \geq \widehat{\underline{L}}_{\min}(B)$ for all $i \in \mathcal{M}$. In practice, however, the minimum can be unstable (sensitive to outliers, label noise, and lucky easy examples within $\mathcal{M}$). A more robust operational choice is to use a small lower quantile,

$$\widehat{\underline{L}}_q(B) \;\equiv\; \mathrm{Quantile}_q\big(\{L_i(\theta_B) : i \in \mathcal{M}\}\big),$$

and interpret the resulting statements as holding for a $1-q$ fraction of $\mathcal{M}$ (or as an approximate floor when the distribution has a sharp lower edge). Either way, the estimation procedure is straightforward: compute $\mathcal{M}$ from π_{ref}, record $L_i(\theta_B)$ (or $\min_{t \leq t_B} L_i(\theta_t)$ if one wants a best-so-far notion), and summarize the lower envelope across mis-ranked points. Repeating across random seeds provides an uncertainty band for $\underline{L}(B)$, which is useful when we later compare predicted ceilings to observed flip rates.

This finite-optimization model is intentionally minimalist: it compresses the training loop into $\mathcal{A}(B)$ and a single floor $\underline{L}(B)$. The benefit is that, when combined with the flip threshold τ_i, it yields a sharp population-level limit on how much fixed-reference DPO can correct among the reference's mis-rankings.

Main population bound: a sharp ceiling from pointwise feasibility. We can now combine the pointwise flip condition with the finite-optimization floor to obtain a population-level limit on how many reference mis-rankings fixed-reference DPO can correct at early stopping. The argument is intentionally simple: each mis-ranked datapoint $i \in \mathcal{M}$ comes with a *required*

per-example loss level $\tau_i = -\log \sigma(\beta c_i)$ that must be met in order to flip the ranking, while the training-and-stopping procedure enforces (in our pessimistic abstraction) a *common* lower bound $\underline{L}(B)$ on what losses are actually reached on $\mathcal{M}$. When the required level τ_i is below the attainable floor $\underline{L}(B)$, that datapoint is infeasible to correct under the given budget and stopping rule.

Formally, fix any $i \in \mathcal{M}$. By the flip threshold characterization (Chen et al., Thm. 4.1; restated in Proposition 1), ranking correctness on i is equivalent to the inequality

$$R_i(\theta_B) = 1 \quad \Longleftrightarrow \quad L_i(\theta_B) \leq \tau_i \equiv -\log \sigma(\beta c_i).$$

Under our finite-optimization assumption, $L_i(\theta_B) \geq \underline{L}(B)$ for all $i \in \mathcal{M}$. Therefore,

$$R_i(\theta_B) = 1 \quad \Longrightarrow \quad \underline{L}(B) \leq \tau_i.$$

This is the key implication: a corrected mis-ranking must lie in the subset of $\mathcal{M}$ whose thresholds are at least as large as the floor. Because $\tau(c) = -\log \sigma(\beta c) = \log(1 + e^{-\beta c})$ is strictly decreasing in c for $c > 0$, we can invert the inequality $\underline{L}(B) \leq \tau(c_i)$ into an equivalent cutoff on the reference log-ratio c_i. Define $c^*(B, \beta)$ as the unique value satisfying

$$-\log \sigma(\beta c^*(B, \beta)) = \underline{L}(B),$$

or equivalently

$$c^*(B, \beta) = \frac{1}{\beta} \operatorname{logit}(e^{-\underline{L}(B)}) = -\frac{1}{\beta} \log\big(e^{\underline{L}(B)} - 1\big).$$

Then for any $i \in \mathcal{M}$,

$$R_i(\theta_B) = 1 \quad \Longrightarrow \quad c_i \leq c^*(B, \beta).$$

Summing this implication over $\mathcal{M}$ yields the correction bound

$$\operatorname{Corr}(B, \beta) = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} R_i(\theta_B) \;\leq\; \frac{1}{|\mathcal{M}|} \left|\{i \in \mathcal{M} : c_i \leq c^*(B, \beta)\}\right|.$$

In large samples, the right-hand side converges to $F_{\mathcal{M}}(c^*(B, \beta))$, the CDF of c restricted to the mis-ranked set. This gives the stated population ceiling:

$$\operatorname{Corr}(B, \beta) \;\leq\; F_{\mathcal{M}}\big(c^*(B, \beta)\big).$$

Uncorrectable tail mass as the operative obstruction. A useful way to read the bound is to emphasize what it says about the *upper tail* of reference disagreement on the mis-ranked set. Rewriting,

$$\operatorname{Corr}(B, \beta) \leq 1 - \mathbb{P}[c > c^*(B, \beta) \mid c > 0].$$

Thus, even if optimization were otherwise perfect, the fraction of mis-ranked points with c above the critical threshold forms an "uncorrectable tail mass" under fixed-reference, early-stopped DPO. These are precisely the datapoints on which the reference assigns exponentially larger odds to the dataset loser y^ℓ than to the dataset winner y^w. For such points, the flip threshold $\tau(c) = \log(1 + e^{-\beta c})$ is extremely small; indeed, for large βc we have $\tau(c) \approx e^{-\beta c}$. If early stopping (or regularization) prevents per-example losses from being driven below $\underline{L}(B)$, then any point requiring $\tau(c) < \underline{L}(B)$ is blocked regardless of what happens on easier datapoints.

This "tail mass" view also clarifies why the bound is not merely a pessimistic artifact of worst-case analysis. The obstruction is structural: under a fixed reference, the DPO loss couples progress on a datapoint to the reference log-odds headwind c_i. When c_i is large and positive, the algorithm must create a sufficiently strong countervailing log-odds shift to overcome the reference preference; doing so requires pushing L_i into a regime that may be inaccessible under the stopping rule. In that sense, the bound formalizes a concrete failure mode: DPO may fit the easy part of $\mathcal{M}$ (small c) while leaving a stubborn subset of high-c mis-rankings unflipped, even though those mis-rankings are precisely where the reference is most wrong relative to the aggregated labels.

Comparative statics: how β and reference quality move the ceiling. The threshold $c^*(B, \beta)$ makes the dependence on β especially transparent. Holding $\underline{L}(B)$ fixed, $c^*(B, \beta)$ scales as $1/\beta$, so increasing β *reduces* the set of mis-ranked points that are even eligible to be corrected:

$$\beta \uparrow \implies c^*(B, \beta) \downarrow \implies F_{\mathcal{M}}(c^*) \downarrow .$$

Operationally, higher β makes DPO behave more like a hard classification of pairwise preferences, which tightens the required per-example loss τ_i on mis-ranked points. Conversely, smaller β relaxes these per-point feasibility constraints, enlarging the feasible region of c values, although at the cost of changing other aspects of training dynamics (including stability and the effective strength of the preference signal).

Reference quality enters only through the conditional distribution of c on $\mathcal{M}$. If the reference is "mildly wrong" on its mis-rankings, then $\mathcal{M}$ may have most of its mass at small positive c, implying a thin tail and a relatively large feasible correction fraction for any given c^*. If, instead, the reference is confidently wrong on a nontrivial subset—a heavy right tail of c on $\mathcal{M}$—then the ceiling can be small even when the overall fraction of mis-rankings $|\mathcal{M}|/n$ is itself small. This is a particularly important safety implication: aggregate win-rate improvements can coexist with a persistent set of high-confidence reference errors that the offline procedure systematically fails to repair under practical stopping rules.

Finally, note what the bound *does not* claim. It is not a guarantee that all datapoints with $c_i \leq c^*$ will be corrected, since optimization can fail for many reasons (gradient interference, limited capacity, or simply insufficient signal). Rather, it identifies a necessary condition for correction under our floor abstraction, and thereby provides an upper envelope: fixed-reference DPO cannot correct more than the mass of mis-ranked points whose reference headwinds are below the critical threshold implied by (B, β) and the induced loss floor. In the next section we discuss how this picture changes once we relax the uniform-floor abstraction and allow heterogeneous learnability or alternative constraints.

4 Extensions (optional but clean)

The preceding ceiling relied on a deliberately blunt abstraction: a *uniform* loss floor over the mis-ranked set, which is a convenient way to model early stopping and finite compute but is not the only operational constraint one might justify from training traces. In this section we sketch three clean extensions that (i) relax uniformity into heterogeneous per-example learnability, (ii) replace pointwise floors with distributional or average-type constraints, and (iii) interpolate the role of the reference term via simple variants that practitioners sometimes implement implicitly.

(i) Heterogeneous learnability as compute allocation; when closed forms fail. A more deployment-faithful view is that different mis-ranked points require different amounts of optimization "effort" before they reach their flip threshold. One stylized way to express this is to endow each $i \in \mathcal{M}$ with a decreasing effort–loss curve, e.g.

$$L_i(e_i) = L_i^0 \exp(-a_i e_i), \qquad e_i \geq 0,$$

where L_i^0 is an initial loss level and a_i is a learnability parameter (capturing gradient signal-to-noise, representation fit, and interference). The pointwise flip condition translates into a minimum effort requirement:

$$R_i = 1 \iff L_i(e_i) \leq \tau_i \iff e_i \geq e_i^* \equiv \frac{1}{a_i} \log \frac{L_i^0}{\tau_i},$$

with the convention that $e_i^* = \infty$ if $L_i^0 \leq \tau_i$ fails to hold in the modeled regime or if τ_i is below a numerical/regularization floor. If we impose a budget constraint $\sum_{i \in \mathcal{M}} e_i \leq B$, then maximizing the number of corrected mis-rankings becomes

$$\max_{e_i \geq 0} \sum_{i \in \mathcal{M}} \mathbf{1}[e_i \geq e_i^*] \quad \text{s.t.} \quad \sum_{i \in \mathcal{M}} e_i \leq B,$$

which is precisely a 0–1 knapsack problem with item "costs" e_i^* and unit values. This reframes the obstruction: even if we discard a uniform floor, the τ_i induced by large c_i still create items with large cost, but now feasibility depends jointly on c_i and (L_i^0, a_i).

Closed forms exist only in special cases. If $a_i \equiv a$ and $L_i^0 \equiv L^0$ are constant across $i \in \mathcal{M}$, then e_i^* is monotone in τ_i, hence monotone in c_i, and the optimal policy is to sort by increasing e_i^* (equivalently, increasing c_i) and take the largest prefix that fits in budget:

$$k^*(B) = \max\Big\{k : \sum_{j=1}^{k} e^*_{(j)} \le B\Big\}, \qquad \mathrm{Corr}(B, \beta) = \frac{k^*(B)}{|\mathcal{M}|}.$$

Outside this homogeneous setting, closed forms generally fail because knapsack is NP-hard in the worst case. This is not merely a mathematical nuisance: it suggests that in realistic training, which points get "fixed" first can depend sensitively on optimization dynamics and representation geometry, not just on the reference log-ratio headwind c_i. From a safety perspective, this creates an additional failure mode beyond the fixed-reference ceiling: even among feasible-to-flip points (moderate c_i), SGD may implicitly allocate effort toward examples with large a_i (easy gradients), leaving behind a residue of hard-but-important mis-rankings that require targeted interventions (curricula, reweighting, or data acquisition).

(ii) Bounds under average-loss or moment constraints (Markov/Cantelli style). The pointwise floor $\forall i \in \mathcal{M} : L_i(\theta_B) \ge \underline{L}(B)$ is a strong pessimistic assumption. A weaker (and often more empirically defensible) constraint is distributional: we may only be willing to assert that the *average* loss over $\mathcal{M}$ at early stopping satisfies

$$\overline{L}_{\mathcal{M}}(\theta_B) \equiv \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} L_i(\theta_B) \;\ge\; \underline{L}_{\mathrm{avg}}(B).$$

By itself, this does not control the corrected fraction: in principle, one could have almost all points below their thresholds while a few points carry arbitrarily large loss, keeping the average high. Consequently, any meaningful ceiling from an average constraint requires an additional boundedness or concentration assumption on $\{L_i(\theta_B)\}_{i \in \mathcal{M}}$.

One clean route (aligned with how many DPO implementations operate) is to assume losses are effectively clipped or otherwise bounded on the training trace, say $0 \le L_i(\theta_B) \le L_{\max}$ for $i \in \mathcal{M}$. Then we can apply a Markov-type argument to the nonnegative variable $X_i \equiv L_{\max} - L_i(\theta_B)$. Since $\mathbb{E}[X] \le L_{\max} - \underline{L}_{\mathrm{avg}}(B)$, Markov yields, for any $t < L_{\max}$,

$$\mathbb{P}\big[L_i(\theta_B) \le t \mid i \in \mathcal{M}\big] = \mathbb{P}\big[X_i \ge L_{\max} - t \mid i \in \mathcal{M}\big] \le \frac{L_{\max} - \underline{L}_{\mathrm{avg}}(B)}{L_{\max} - t}.$$

To translate this into a correction ceiling, we upper bound the event $\{R_i(\theta_B) = 1\}$ by $\{L_i(\theta_B) \le \tau(c_i)\}$ and average over the conditional law of c on $\mathcal{M}$:

$$\mathrm{Corr}(B,\beta) = \mathbb{E}\big[\mathbf{1}[R=1] \mid c>0\big] \le \mathbb{E}\big[\mathbf{1}[L \le \tau(c)] \mid c>0\big] \le \mathbb{E}\left[\min\left\{1, \frac{L_{\max} - \underline{L}_{\mathrm{avg}}(B)}{L_{\max} - \tau(c)}\right\} \,\middle|\, c>0\right].$$

Unlike the sharp cutoff obtained under a pointwise floor, this produces a "soft" ceiling that depends on the whole distribution of $\tau(c)$ and on the boundedness proxy $L_{\max}$. If one can estimate not just the mean but also variance of losses on $\mathcal{M}$, one can swap Markov for one-sided Chebyshev/Cantelli inequalities to obtain tighter ceilings; the main conceptual point is that *moment information about the loss distribution* can substitute for a uniform floor, but some form of tail control is necessary.

(iii) Variants: γ-scaled reference terms and related knobs. Practitioners sometimes temper the influence of the reference model, either explicitly (e.g., scaling reference logits) or implicitly (e.g., via length normalization, truncation, or mixture references). A minimal abstraction is to introduce a scalar $\gamma \ge 0$ multiplying the reference log-ratio term, yielding a modified per-point loss

$$L_i^{(\gamma)}(\theta) = -\log\sigma\left(\beta\left(\log\frac{\pi_\theta(y_i^w \mid x_i)}{\pi_\theta(y_i^\ell \mid x_i)} + \gamma c_i\right)\right).$$

In this variant, the flip threshold becomes $\tau_i^{(\gamma)} = -\log\sigma(\beta\gamma c_i)$. Holding the same early-stopping floor abstraction, the critical threshold rescales as

$$c^{*,(\gamma)}(B,\beta) = \frac{1}{\beta\gamma}\,\mathrm{logit}(e^{-\underline{L}(B)}),$$

so decreasing γ (weakening the reference headwind) expands the feasible set of c values and can strictly increase the attainable correction ceiling. The governance-relevant tradeoff is that $\gamma < 1$ also weakens anchoring to π_{ref}, plausibly increasing divergence and capability shift, and may amplify exploitation of label noise. Thus γ functions as a policy knob: it buys correction headroom by partially relaxing the very constraint that makes fixed-reference DPO stable and predictable.

These extensions give us a menu of testable refinements: we can ask whether observed flip patterns are better explained by a sharp c-cutoff (uniform floor), by heterogeneous costs e_i^* (knapsack-like behavior), or by soft ceilings derived from bounded average losses. This sets up the empirical validation plan that follows.

(iv) Empirical validation plan: estimating c, L_i, flips, and testing the ceiling. The theory above is only as useful as our ability to operationalize its objects from standard training artifacts (log-probabilities, per-example losses, and checkpoints). Our empirical goal is therefore narrowly

scoped: given a fixed preference dataset D, a fixed reference π_{ref}, and a family of DPO runs indexed by (β, B) and random seed, we test whether the observed correction rate on $\mathcal{M}$ is upper-bounded (up to estimation error) by the predicted ceiling $\widehat{F}_{\mathcal{M}}(\widehat{c}^*)$ implied by the early-stopping loss floor. When the bound appears loose, we treat this as informative about slack in the floor assumption; when the bound is violated, we treat it as evidence of a modeling mismatch (e.g., an effectively changing reference, an implementation detail that alters c_i, or a flip definition not aligned with Chen et al.'s threshold).

Step 1: compute reference log-ratios and the mis-ranked set. For each datapoint i, we compute

$$c_i \equiv \log \pi_{\text{ref}}(y_i^\ell \mid x_i) - \log \pi_{\text{ref}}(y_i^w \mid x_i),$$

using teacher-forced log-likelihoods under the same tokenizer and truncation policy used for DPO. We then define $\mathcal{M} = \{i : c_i > 0\}$. Two practical details matter. First, sequence-length differences can induce systematic shifts in $\log \pi(\cdot \mid x)$; accordingly, we pre-register two variants: the *sequence-sum* log-probability above, and a *length-normalized* alternative $c_i^{\text{len}} \equiv \frac{1}{|y_i^\ell|} \log \pi_{\text{ref}}(y_i^\ell \mid x_i) - \frac{1}{|y_i^w|} \log \pi_{\text{ref}}(y_i^w \mid x_i)$. Second, truncation can flip the sign of c_i for long completions; we therefore report the fraction of pairs whose sign changes under plausible truncation windows and treat high sensitivity as an exclusion/stratification criterion (since it confounds what it means to be "mis-ranked by the reference").

Step 2: record per-example DPO losses along the training trace. For each run and each saved checkpoint θ_t (including the selected early-stopped checkpoint θ_B), we compute the per-example DPO loss

$$L_i(\theta_t) = -\log \sigma\left(\beta\left(\log \frac{\pi_{\theta_t}(y_i^w \mid x_i)}{\pi_{\theta_t}(y_i^\ell \mid x_i)} + c_i\right)\right),$$

either exactly (offline evaluation over D) or approximately (via cached forward passes if training logs already store the needed log-probabilities). We focus attention on the restriction $\{L_i(\theta_B) : i \in \mathcal{M}\}$, since the ceiling is controlled by what happens on mis-ranked points. In addition to raw values, we log summary statistics stratified by c_i quantiles, because a core mechanistic claim is that large positive c_i induces very small flip thresholds τ_i, hence should correlate with persistently large losses under early stopping.

Step 3: define and measure "flip" events in a numerically stable way. The ranking-correctness event is

$$R_i(\theta_B) = \mathbf{1}\Big[\pi_{\theta_B}(y_i^w \mid x_i) \geq \pi_{\theta_B}(y_i^\ell \mid x_i)\Big] = \mathbf{1}\left[\log \frac{\pi_{\theta_B}(y_i^w \mid x_i)}{\pi_{\theta_B}(y_i^\ell \mid x_i)} \geq 0\right].$$

In finite precision, near-ties are common; we therefore report both (a) the inclusive definition above and (b) a margin-based definition $R_i^{(\varepsilon)} = \mathbf{1}[\log \frac{\pi_{\theta_B}(y_i^w|x_i)}{\pi_{\theta_B}(y_i^\ell|x_i)} \geq$

$\varepsilon]$ for a small $\varepsilon > 0$. We also report a tie-aware score that assigns $1/2$ when the log-odds magnitude is below a tolerance band. These variants are not cosmetic: if most "flips" occur at vanishing margins, then the governance-relevant conclusion is weaker (the model may be indifferent rather than reliably aligned), and the ceiling should be interpreted in terms of margins rather than strict orderings.

Step 4: estimate the early-stopping loss floor and plug in the predicted ceiling. Because $\underline{L}(B)$ is an abstraction, we treat it as an *estimand* from training traces rather than a known constant. Our default estimator is a robust lower-quantile of losses on $\mathcal{M}$,

$$\widehat{\underline{L}}(B) \equiv \mathrm{Quantile}_q\left(\{L_i(\theta_B) : i \in \mathcal{M}\}\right),$$

with $q \in \{0.01, 0.05, 0.10\}$ as a sensitivity parameter. This intentionally avoids using $\min_i L_i(\theta_B)$, which is brittle to outliers and logging noise, and it corresponds to interpreting the "floor" as what early stopping prevents *most* mis-ranked points from undercutting. We then compute

$$\widehat{c}^*(B,\beta) = \frac{1}{\beta}\,\mathrm{logit}\Big(e^{-\widehat{\underline{L}}(B)}\Big), \qquad \widehat{\mathrm{Corr}}_{\max}(B,\beta) = \widehat{F}_{\mathcal{M}}(\widehat{c}^*(B,\beta)),$$

where $\widehat{F}_{\mathcal{M}}$ is the empirical CDF of $\{c_i : i \in \mathcal{M}\}$. We report bootstrap confidence intervals that jointly resample datapoints (to reflect $\widehat{F}_{\mathcal{M}}$ uncertainty) and seeds (to reflect optimization variability). The predicted $\widehat{\mathrm{Corr}}_{\max}$ is then compared to the observed $\widehat{\mathrm{Corr}}(B,\beta) = \frac{1}{|\mathcal{M}|}\sum_{i\in\mathcal{M}} R_i(\theta_B)$.

Step 5: test the "upper-envelope" prediction across datasets, β, and budgets. We run a grid over β and compute budgets B (or, operationally, over early-stopping checkpoints selected by a fixed rule). For each setting, we plot $\widehat{\mathrm{Corr}}(B,\beta)$ against $\widehat{\mathrm{Corr}}_{\max}(B,\beta)$ and test whether $\widehat{\mathrm{Corr}}$ concentrates below the 45° line. The cross-dataset prediction is that datasets with heavier upper tails of c on $\mathcal{M}$ (i.e., more severe reference mis-rankings) exhibit lower attainable correction at comparable (β, B), even when average win-rate improvements look similar. Mechanistically, we also expect a sharp transition in flip incidence around $\widehat{c}^*$: plotting the empirical flip rate as a function of c (e.g., in bins) should reveal a decline consistent with the monotonicity of $\tau(c)$.

Robustness and falsification checks. We pre-register several checks designed to separate genuine violations from definitional artifacts:

1. **Length normalization and truncation:** repeat the full pipeline with c_i^{len} and with multiple truncation windows; report how $\widehat{F}_{\mathcal{M}}$ and $\widehat{c}^*$ shift.

2. **Alternative flip definitions:** replace R_i with a probabilistic proxy, e.g., $\mathbf{1}[\sigma(\log \frac{\pi_{\theta_B}(y_i^w|x_i)}{\pi_{\theta_B}(y_i^\ell|x_i)}) \geq 1/2]$ (equivalent) versus a higher threshold

$p \geq 0.6$ (margin-sensitive); test whether the ceiling becomes tighter for larger margins, as the threshold picture would suggest.

3. **Ties and label ambiguity:** isolate datapoints with small rater margins (if available) or near-equal reference probabilities $|c_i| \approx 0$; these are precisely where majority labels and model likelihoods are least stable, so they can dominate apparent "violations".

4. **Implementation drift from a fixed reference:** verify that training truly uses π_{ref} as specified (no moving-average reference, no implicit mixture, no logit scaling). If a γ-like effect is present, recompute $c^{*,(\gamma)}$ and re-evaluate.

Interpreting outcomes is straightforward: persistent slack suggests our floor estimate is conservative or that optimization allocates effort heterogeneously; systematic violations, especially concentrated at large c, point toward an effectively changing reference or a mismatch between the theoretical loss and the implemented objective. Either way, the validation pipeline turns an otherwise qualitative claim ("DPO rarely flips mis-rankings") into a quantitative, dataset-conditioned diagnostic that practitioners can run before attributing failures to "insufficient training."

8. Implications for 2026 alignment governance: when "more preference tuning" is futile, and what to do instead. A practical governance question in 2026 is no longer whether preference tuning *can* improve a model in aggregate, but whether additional rounds of DPO on a fixed dataset and fixed reference can be expected to correct the specific mis-rankings that matter for safety and policy compliance. The bound above gives us a concrete answer in terms that are auditable from standard artifacts: if the mis-ranked set $\mathcal{M}$ has substantial mass at large positive reference log-ratio c, then there exists an *uncorrectable tail* $\{i \in \mathcal{M} : c_i > c^*(B, \beta)\}$ that cannot be flipped by any early-stopped solution whose per-point losses remain above $\underline{L}(B)$. In that regime, "more preference tuning" (in the sense of rerunning the same protocol with slightly more steps, or sweeping seeds) predictably produces diminishing returns: it may improve easy or moderate points, but it does not buy back the hard tail where the reference is confidently wrong.

Futility as a measurable stopping condition (and why it matters for oversight). Operationally, futility is not a philosophical claim; it is a thresholded diagnostic. Given a planned run (B, β), practitioners can precompute $\{c_i : i \in \mathcal{M}\}$, estimate a plausible loss floor $\underline{L}(B)$ from prior runs, and thereby obtain a predicted ceiling $F_{\mathcal{M}}(c^*(B, \beta))$. If that ceiling is, say, 20% while a safety case demands correcting 60% of reference mis-rankings in a policy-critical slice (e.g., high-severity refusals), then the correct governance conclusion is that the current training protocol is *structurally incapable*

of meeting the target. This is exactly the kind of determination that internal safety reviews and external auditors should prefer over informal claims like "we trained longer" or "we tried more seeds": the ceiling translates those claims into an empirically testable prediction about what improvement is even feasible without changing the reference, the data, or the algorithm.

Interpreting β as inertia (or an adjustment cost) rather than a mere hyperparameter. In DPO, β scales the preference odds, but in the presence of a fixed reference it also governs how hard it is to override the reference on mis-ranked points. For $i \in \mathcal{M}$, the flip threshold is $\tau_i = -\log \sigma(\beta c_i)$, which shrinks as β increases (holding $c_i > 0$ fixed). Thus larger β implements a form of *inertia*: it makes corrections feasible only when the optimizer can drive the loss extremely low, which early stopping and finite compute often prevent. From a governance perspective, β is therefore interpretable as an *adjustment cost* or "status quo bias" against changing the reference model's ordering in precisely those cases where the reference disagrees with the aggregated labels. This makes explicit a tradeoff that is otherwise implicit in deployment decisions: choosing a large β can protect against uncontrolled drift away from π_{ref}, but it also locks in some portion of the reference's failures; choosing a smaller β relaxes that lock-in, but typically increases variance and the risk of capability degradation or unintended generalization. Treating β as a governance knob suggests it should be set with a *documented* rationale tied to (i) the tail behavior of c on $\mathcal{M}$, and (ii) the organization's tolerance for divergence from π_{ref} in exchange for correcting known mis-rankings.

When to invest in better references versus more compute. The bound clarifies when the marginal dollar should go to compute versus reference improvement. If $\underline{L}(B)$ is already low (training is effective on easy points) but $F_{\mathcal{M}}(c^*(B,\beta))$ remains small because $c^*(B,\beta)$ sits far left of the c-distribution's upper tail, then more compute buys little: the bottleneck is that the reference assigns overwhelming odds against the majority-preferred completion on a subset of $\mathcal{M}$. In that case, the most cost-effective intervention is to *shift the c-distribution left* by improving π_{ref}: stronger pretraining, better instruction tuning, domain-specific calibration, or simply a reference trained on a broader preference dataset. In contrast, if the c-tail is mild but the estimated $\underline{L}(B)$ is high (training under the current budget fails to push losses down even on moderate points), then additional budget or better optimization (batching, curriculum, per-example weighting, more stable implementations) is likely to increase $c^*(B,\beta)$ and therefore raise the attainable correction mass.

On-policy refresh and iterative protocols: how they evade the fixed-reference ceiling (and what can go wrong). A central implication for training protocol design is that the ceiling is a *fixed-reference* phenomenon. If we refresh the reference over iterations (iterative DPO, on-policy RLHF variants, or periodically re-baselining $\pi_{\text{ref}} \leftarrow \pi_\theta$), then difficult points may become feasible because their effective c shrinks as the reference moves. This is a legitimate way to escape the uncorrectable tail, and it explains why practitioners sometimes observe improvements that a single-shot fixed-reference analysis would deem impossible. However, from a safety standpoint, this maneuver changes the object being governed: the "reference" ceases to be a stable anchor, and the organization must monitor for feedback-loop failures (reward hacking, over-optimization of rater quirks, or distributional drift in refusal behavior). Governance here should treat iterative refresh not as "more of the same tuning" but as a *protocol change* requiring additional safeguards: hold-out evaluations, adversarial red-teaming, and explicit constraints on divergence or on policy-critical behaviors.

Concrete diagnostics we recommend practitioners pre-register. To make these implications usable, we can standardize a small set of diagnostics that translate the theory into actionable go/no-go decisions:

1. **Tail mass report:** for each dataset slice of interest (especially safety-critical prompts), report $\widehat{\mathbb{P}}[c > t \mid c > 0]$ for a grid of t, not just the mean of c. The feasibility bottleneck is tail-driven.

2. **Ceiling card per run:** for each (B, β), report $\widehat{\underline{L}}(B)$, $\widehat{c}^*(B, \beta)$, and $\widehat{F}_{\mathcal{M}}(\widehat{c}^*)$ alongside the observed $\widehat{\text{Corr}}(B, \beta)$. This makes "we trained longer" falsifiable.

3. **Flip-by-c curve:** bin $\mathcal{M}$ by c and plot empirical flip rates. A sharp drop around $\widehat{c}^*$ supports the mechanism; a flat curve suggests either measurement error or a non-DPO effect.

4. **Uncorrectable set inspection:** sample datapoints with $c_i \gg \widehat{c}^*$ and review them qualitatively. If they correspond to policy-critical failures, governance should mandate reference improvement or protocol change rather than incremental tuning.

5. **β-sweep as an "inertia test":** run a small sweep over β at fixed budget to empirically map the correction–divergence tradeoff; require that chosen β sits on an explicit frontier rather than being inherited from default recipes.

Protocol-level recommendations for governance decisions. Putting these pieces together, we can articulate a simple decision rule: if the predicted

ceiling is low due to heavy c-tails, invest in (i) a better reference, (ii) data that directly targets the hard tail (more informative comparisons, adjudication, or task decomposition), or (iii) iterative/on-policy refresh with strengthened monitoring. If the ceiling is low due to a high estimated $\underline{L}(B)$, invest in optimization and compute. Crucially, these choices correspond to different risk postures: improving π_{ref} tends to preserve anchoring while reducing mis-rank severity; iterative refresh can correct more but increases the need for governance around drift. This framing sets up the next section: our bound is informative precisely because it is limited, and we should be explicit about what it does not capture when translating it into claims about win rates, welfare, and deployment safety.

9. Discussion and limitations: from a correction ceiling to welfare claims. The object our analysis bounds is $\text{Corr}(B, \beta)$, a *ranking-correctness* statistic on the mis-ranked set $\mathcal{M}$ induced by the fixed reference π_{ref}. This is a natural proxy when the deployment goal is "make the model agree with aggregated preferences on the particular pairs where the reference was wrong." But it is not, by itself, a guarantee about on-policy win rate or downstream welfare. In particular, an increase in (on-policy) win rate can occur without correcting many elements of $\mathcal{M}$ if training mostly improves already-correct comparisons (e.g., by sharpening margins where $c_i < 0$ and the reference already agrees with the majority). Conversely, one can correct a sizeable fraction of $\mathcal{M}$ and still see little win-rate improvement if the corrected pairs are rare under the deployment distribution $\mathcal{X}$, or if the evaluation distribution differs from the curated preference dataset D. In our principal objective $U_{\text{principal}}(\beta, B)$, $\text{Corr}(B, \beta)$ is therefore best viewed as a *verifiable component* of welfare—an intermediate quantity that can be audited from training artifacts—rather than the full welfare signal.

Correction is not divergence, and divergence is not safety. A second limitation is conceptual: our ceiling is compatible with two different failure modes that are often conflated in practice. First, one may fail to correct $\mathcal{M}$ because the loss cannot be pushed below the required thresholds τ_i at finite compute (our mechanism). Second, one may be able to correct more of $\mathcal{M}$ by choosing smaller β or training longer, but doing so may incur unacceptable divergence from π_{ref} and therefore degrade capabilities or violate other constraints. The bound speaks to the first phenomenon, while the principal's regularization term $\text{Deg}(\pi_{\theta_B}, \pi_{\text{ref}})$ speaks to the second. Neither alone determines safety. In deployment, what typically matters is a *risk-weighted* objective: a small uncorrectable tail can still be unacceptable if it concentrates high-severity failures. A straightforward extension is to replace $\text{Corr}(B, \beta)$ with a weighted correction rate $\text{Corr}_w(B, \beta) = \sum_{i \in \mathcal{M}} w_i R_i(\theta_B) / \sum_{i \in \mathcal{M}} w_i$, in which case the same logic yields an upper bound in terms of the weighted

CDF of c over $\mathcal{M}$. The governance implication is that tail mass should be reported not only in aggregate, but also within slices defined by severity, policy category, or misuse potential.

What the ceiling does *not* capture mechanistically. The ceiling is deliberately narrow: it isolates a fixed-reference DPO phenomenon under an early-stopping floor. It does not capture (i) improvements that come from changing the data distribution (e.g., collecting comparisons that make the hard tail easier), (ii) improvements that come from changing the reference (iterative baselining), or (iii) improvements that come from changing the objective (e.g., adding explicit constraints or auxiliary losses). It also does not model generalization, either beneficial or harmful: in practice, a DPO update that fails to flip a particular pair (x_i, y_i^w, y_i^ℓ) may still generalize and reduce analogous mis-rankings elsewhere, and conversely may overfit to spurious artifacts that inflate $\mathrm{Corr}(B, \beta)$ on D while worsening behavior off-distribution. Finally, we have treated the aggregated label $y_i^w \succ y_i^\ell$ as ground truth. When rater disagreement is substantial, the relevant target may be probabilistic, and the most appropriate metric may be expected regret under a rater model rather than a deterministic correction indicator.

On the floor assumption: when it is reasonable, and how to relax it. The finite-optimization floor $\forall i \in \mathcal{M} : L_i(\theta_B) \geq \underline{L}(B)$ is a stylized way to encode the observation that early-stopped solutions do not drive every per-example loss arbitrarily low. It is most plausible when (a) we stop based on a global criterion (validation loss, wall-clock, or stability), (b) gradients are noisy and optimization is approximate, and (c) $\mathcal{M}$ contains hard outliers whose gradients are either rare or conflict with more common points. That said, a *uniform* deterministic floor is stronger than needed. Two relaxations are natural.

First, we can assume a *probabilistic floor* such as

$$\mathbb{P}\big[L_i(\theta_B) \geq \underline{L}(B) \mid i \in \mathcal{M}\big] \geq 1 - \delta,$$

which yields a slackened ceiling $\mathrm{Corr}(B, \beta) \leq F_{\mathcal{M}}(c^*(B, \beta)) + \delta$. This matches empirical reality: a small fraction of points may be fit extremely well, but most are not. Second, we can allow *heterogeneous floors* $\underline{L}_i(B)$ that depend on prompt length, rarity, or gradient signal-to-noise. Then

$$R_i(\theta_B) = 1 \;\Rightarrow\; \underline{L}_i(B) \leq \tau_i,$$

and the predicted attainable correction becomes a (computable) expectation of the form $\mathbb{E}[\mathbf{1}[\underline{L}_i(B) \leq \tau(c_i)] \mid i \in \mathcal{M}]$. This generalization connects directly to optimization modeling: if we can predict which points are intrinsically hard to fit, we can distinguish "uncorrectable due to reference confidence" (large c_i) from "uncorrectable due to optimization difficulty" (large $\underline{L}_i(B)$).

Estimating $\underline{L}(B)$ without circularity. A practical worry is that $\underline{L}(B)$ may look like an ex post quantity: we only know it after training. For governance, we want ex ante predictions that can be pre-registered. One approach is to fit an empirical learning curve for a robust statistic of losses on $\mathcal{M}$, e.g., the α-quantile of $\{L_i(\theta_t) : i \in \mathcal{M}\}$ over training time t, and extrapolate it to the planned budget B. Another is to use cross-run reproducibility: if the lower tail of per-point losses across seeds is stable, we can treat its plateau as an operational floor. The crucial methodological point is that the ceiling becomes meaningful when it is treated as a *forecast* that can be falsified: if observed correction persistently exceeds the forecast, we should update our model (e.g., the floor is not binding, or the effective reference has changed); if observed correction persistently falls short, we should suspect additional bottlenecks (data noise, mis-specified $\mathcal{M}$, or optimization pathologies).

Future directions: dynamic references as a controlled escape hatch. Iterative refresh protocols evade the fixed-reference ceiling precisely by changing the c-distribution for the remaining hard points. This suggests a principled design question: can we treat reference updates as a controlled process that preserves anchoring while shrinking the uncorrectable tail? One direction is to formalize a *dynamic reference schedule* $\pi_{\text{ref}}^{(k)}$ with constraints such as $\mathrm{KL}(\pi_{\text{ref}}^{(k+1)} \| \pi_{\text{ref}}^{(k)}) \leq \epsilon$ and explicit auditing checkpoints between updates. In such a protocol, we can re-apply the ceiling analysis *per iteration* to predict how much of the remaining tail can be corrected before the next refresh, and we can interpret failures as either "the tail is too heavy" or "the refresh step is too conservative." The open problem is to identify conditions under which these dynamics converge to a policy that both satisfies preferences and remains within acceptable divergence and safety constraints, rather than entering a feedback loop that optimizes rater idiosyncrasies.

Future directions: market-like auditing and adversarial procurement of signal. Finally, the ceiling suggests a new kind of auditability: because c_i and empirical loss traces are standard artifacts, third parties can often compute (or approximate) the predicted correction envelope without access to proprietary gradients or full training code. This creates room for "market-like" oversight mechanisms. For example, an internal evaluation team (or an external auditor under NDA) could run a prediction-market-style process over $\widehat{\mathrm{Corr}}(B, \beta)$ for a planned run, with participants rewarded for accurate forecasts based on $\{c_i\}$ and prior $\underline{L}(B)$ estimates. More substantively, an organization can *procure* information by paying for comparisons targeted at the hard tail: identify datapoints with c_i above the current $c^*(B, \beta)$, then solicit higher-quality adjudication (expert raters, richer rubrics, or decomposed subtasks) to either (i) confirm that the labels are correct and therefore demand protocol changes, or (ii) reveal that the aggregate label was noisy

and the point should not be treated as a safety-relevant mis-ranking. In this sense, the ceiling is not only a limitation; it is also a guide for where additional oversight effort is most valuable.

Bottom line. Our bound is intentionally modest: it does not solve alignment, and it does not certify safety. What it does is turn a familiar empirical observation—that fixed-reference preference tuning often fails to overturn confident reference mistakes under early stopping—into a quantitatively testable constraint tied to measurable tail behavior. The main limitation is also the main opportunity: once we can diagnose futility in advance, we can make governance decisions that change the right lever (reference quality, data targeting, or protocol class) rather than repeating a training recipe whose improvement path is, in a precise sense, already exhausted.