# A UNIFIED RANDOM-UTILITY VIEW OF HETEROGENEITY IN DIRECT PREFERENCE OPTIMIZATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Direct Preference Optimization (DPO) assumes a single comparison scale, although in practice some response pairs are easy to distinguish while others remain ambiguous at similar utility gaps. Recent DPO-like variants increasingly introduce pair-adaptive temperatures, margins, confidence signals, or latent structure to address heterogeneity, yet similar language is often used for methods that model different forms of heterogeneity at different stages of preference modeling. To disentangle these differences, we present a unified random-utility view of heterogeneity in DPO and use it to organize DPO-like methods into a branch-level taxonomy. This taxonomy isolates a statistically distinct DPO-side error-layer branch, comparison-specific observation-scale heterogeneity, in which homoscedastic DPO is semantically misspecified: to match correct preference probabilities, its score must absorb comparison-specific observation scale, converging to a scale-distorted target rather than the latent utility gap itself. Its minimal instantiation is Heteroscedastic Direct Preference Optimization (H-DPO), a heteroscedastic-logit generalization of DPO that preserves the standard utility-gap parameterization while introducing comparison-specific scale in the logit. Under restricted smooth families, the corresponding score-scale decomposition is locally identifiable up to global scaling from binary preference data. Synthetic experiments confirm the predicted score distortion, and real-world benchmarks support the comparison-scale formulation and its preference-discrimination benefits.

## 1 INTRODUCTION

Direct Preference Optimization (DPO) (Rafailov et al., 2023) is a widely used method for preference-based post-training of large language models (LLMs) (Azar et al., 2024; Ethayarajh et al., 2024; Meng et al., 2024). Standard DPO assumes a single global comparison scale: once the latent preference gap between two responses is fixed, the sharpness of the observed binary preference is fixed as well. Real preference pairs, however, are often not equally clear-cut: some are nearly deterministic, while others remain ambiguous even for similar underlying preference gaps. Recent DPO-like methods increasingly introduce pair-adaptive temperatures, margins, confidence signals, semantic strengths, or latent user structure (Houliston et al., 2024; Pokharel et al., 2025; Imai et al., 2026; Wu et al., 2024; Chidambaram et al., 2025; Son et al., 2025). But these methods need not model the same statistical object. Some modify latent utility, some enrich supervision, some model temporal variation, and many change only the training objective. As a result, similar language (such as uncertainty, confidence, preference strength, or dispersion) is often used for different forms of heterogeneity with different supervision requirements and different inferential consequences under binary pairwise data. By 2026, this has become a practical problem rather than a terminological one: the DPO literature contains many pair-adaptive variants, but they often modify different statistical objects. Without separating these branches first, it is unclear what is being modeled, what supervision is required, and what binary pairwise data can actually identify.

To disentangle these differences, we develop a unified random-utility view of heterogeneity in direct preference optimization and use it to distinguish utility-side, observation-scale, outcome-side, temporal, and loss-layer adaptations. Within this taxonomy, we isolate a DPO-side error-layer branch: **comparison-specific observation-scale heterogeneity**. In this branch, a shared latent utility system is observed through comparison-specific observation scales, so that the same utility gap can induce different preference sharpness across response pairs. Its key consequence is semantic rather



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.

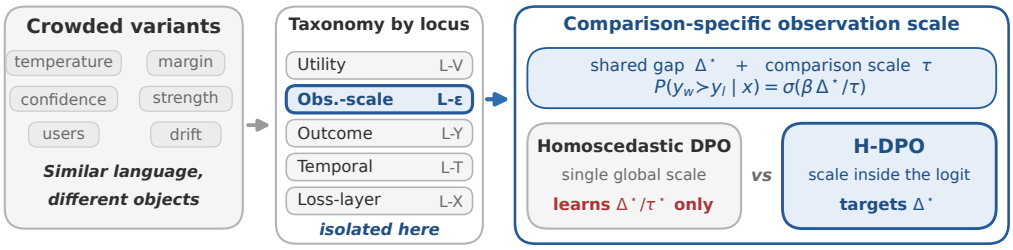



Figure 1: **A branch-level view of heterogeneity in DPO.** Many DPO-like methods introduce pair-adaptive quantities under overlapping language, but these quantities can enter different parts of the preference pipeline. Our taxonomy separates utility-side, observation-scale, outcome-side, temporal, and loss-layer adaptations, and isolates a distinct DPO-side branch of comparison-specific observation-scale heterogeneity. In this branch, the same latent utility gap can induce different preference sharpness across pairs. As a result, homoscedastic DPO can match preference probabilities only by learning a scale-distorted score, whereas H-DPO models comparison-specific scale directly as part of the latent observation channel.

than merely predictive: comparison-specific ambiguity changes what homoscedastic DPO learns. If the true process contains comparison-specific observation scale, then a homoscedastic DPO score can match observed preference probabilities only by absorbing that scale variation into the score itself. This is the paper's central claim: comparison-specific observation scale changes not only how preferences should be fit, but also what quantity homoscedastic DPO learns.

As a minimal DPO-side instantiation of this branch, we study **Heteroscedastic Direct Preference Optimization (H-DPO)**, a heteroscedastic-logit generalization of DPO that preserves the standard utility-gap parameterization and relaxes only the homoscedastic comparison-scale assumption by introducing comparison-specific scale inside the logit. Throughout this paper H-DPO refers to this heteroscedastic model and not to the entropy-controlled H-DPO of Omura et al. (2024). It is a minimal instantiation of this branch rather than another pair-adaptive objective. Our contribution is therefore not heteroscedastic logit per se, but its identification as a distinct DPO-side branch of comparison-specific observation-scale heterogeneity and its semantic consequence for DPO.

The contributions are: (i) a **unified random-utility taxonomy** of DPO-like methods that distinguishes utility-side, observation-scale, outcome-side, temporal, and loss-layer heterogeneity; (ii) isolation of a DPO-side error-layer branch, **comparison-specific observation-scale heterogeneity**, together with its minimal instantiation as **H-DPO**; (iii) theoretical results showing that homoscedastic DPO is **semantically misspecified** for this branch, that positive loss-layer reweighting does not remove the resulting target distortion, and that under restricted smooth families the heteroscedastic branch is **locally identifiable up to global scaling**; and (iv) empirical validation on synthetic and real-world benchmarks confirming the predicted score distortion, supporting the inside-logit comparison-scale mechanism, and showing that the learned scale exhibits interpretable comparison-specific behavior in real data.

## 2 A Unified Random-Utility View of DPO

Figure 1 summarizes our branch-level view of heterogeneity in DPO. We place standard DPO in a common random-utility template, distinguish the *source* of heterogeneity from the *locus* at which it is modeled, and organize DPO-like methods into a taxonomy that isolates the comparison-specific observation-scale branch studied here.

### 2.1 Preference Optimization as a Random-Utility Model

We begin with a common random-utility template. Let $x$ denote a prompt and $y$ a candidate response. A response is assigned latent utility

$$U(x, y) = V(x, y) + \varepsilon(x, y), \tag{1}$$

where $V(x, y)$ is a systematic utility component and $\varepsilon(x, y)$ is a stochastic utility component. Given two responses $y_1$ and $y_2$, observed supervision is generated by comparing their latent utilities through an observation rule

$$Y \sim \mathcal{O}(U(x, y_1), U(x, y_2)), \tag{2}$$

where $Y$ may be a binary choice, a tie, a graded label, or a richer ranking signal depending on the form of $\mathcal{O}$. This template provides a common language for DPO-like methods. Different methods may modify the systematic utility $V$, the stochastic utility component $\varepsilon$, the observation rule $\mathcal{O}$, the stationarity of the process, or only the training objective. Standard DPO corresponds to the binary, homoscedastic branch of this template. It parameterizes the utility gap through policy-relative log-probability ratios against a reference policy and uses a single global comparison scale, yielding the Bradley–Terry / conditional-logit baseline used throughout the paper.

### 2.2 Source and Locus of Heterogeneity

We distinguish two questions: *why* preferences vary and *where* the model encodes that variation. We call the first the **source** of heterogeneity and the second its **locus**. This distinction matters because similar informal terms (such as uncertainty, ambiguity, preference strength, or dispersion) need not refer to the same statistical object. For our purposes, relevant sources include user-level variation, comparison-specific ambiguity, enriched supervision, temporal drift, and heuristic adaptation.

Correspondingly, heterogeneity can enter at several loci: the **utility layer** (L-V, modifying $V$); the **error layer** (L-$\varepsilon$, modifying comparison noise or observation scale); the **outcome layer** (L-Y, modifying $\mathcal{O}$); the **temporal layer** (L-T, explicit non-stationarity); and the **loss layer** (L-X, pair-dependent temperatures, weights, or margins introduced only in the objective). Source and locus are not in one-to-one correspondence: similar preference probabilities can still reflect different latent structure, supervision requirements, and identifiability properties. The practically important distinction for this paper is between L-$\varepsilon$ and L-X. A method belongs to the error layer when the comparison-specific quantity is modeled inside the *conditional preference model* itself as part of the latent observation channel, thereby changing how latent utility differences are translated into observed preferences. By contrast, loss-layer methods introduce pair-adaptive quantities only as optimization surrogates or external controls. The distinction is therefore not merely whether a scalar appears inside the logit, but whether it is modeled as part of the latent observation channel.

### 2.3 A Branch-Level Taxonomy of DPO-like Methods

Table 1 summarizes this design space at the branch level: where heterogeneity enters, what supervision is required by construction, and what binary pairwise data typically support in that branch. It is a statistical map of the main neighboring families rather than an exhaustive survey. In particular, methods that look similarly pair-adaptive may differ sharply in whether they define latent generative structure, require richer labels, or are identifiable from binary comparisons.

### 2.4 A DPO-side Error-Layer Branch

The taxonomy highlights a statistically distinct DPO-side branch: **comparison-specific observation-scale heterogeneity**. It arises even when the score class is sufficiently flexible. The issue is not whether the score can fit the data, but what the score means when the true process contains comparison-specific observation scale and the fitted model assumes a homoscedastic observation channel. In this branch, a shared latent utility system is observed through comparison-specific observation scales, so that the same utility gap can induce different preference sharpness across response pairs. Our focus is the specific error-layer case in which comparison-specific observation scale is part of the conditional preference model itself under binary pairwise supervision. By construction, this branch requires neither richer labels nor user identities, but it does introduce a genuine confounding between utility magnitude and local observation scale. Section 3 instantiates this branch as Heteroscedastic Direct Preference Optimization (H-DPO).

Table 1: Branch-level taxonomy. Identifiability entries summarize typical support from binary pairwise data, not per-method theorems.

| Method family | Source | Locus | Supervision | Identifiability |
|---|---|---|---|---|
| Homoscedastic DPO variants (Rafailov et al., 2023; Azar et al., 2024; Meng et al., 2024; Park et al., 2024) | none | none | binary pairs | yes, under standard normalization |
| Utility mixtures (Imai et al., 2026; Chidambaram et al., 2025) | user-level coefficients / latent types | L-V | binary + user IDs or ternary labels | partial; often not from binary alone |
| Outcome enrichment (Liu et al., 2024; Furuta et al., 2024; Kim et al., 2024) | ties / soft or graded labels | L-Y | ties / soft labels / ratings | identifiable with richer labels |
| Temporal heterogeneity (Son et al., 2025) | preference drift | L-T | binary + timestamps | conditional on temporal structure |
| Objective-/process-level adaptation (Hong et al., 2024; Wu et al., 2024; Wu et al., 2025; Lee et al., 2025; He et al., 2026; Rho, 2025) | pair/batch confidence, semantic strength, margin, or temperature | L-X | binary; some auxiliary signals | no latent structure by construction |
| Prompt-level dispersion / ranking-family formulations (Chen et al., 2025) | prompt-level dispersion | prompt-conditioned scaling | binary pairs | model-dependent |
| RM-side error-scale counterparts (Liu et al., 2025; Sun et al., 2025; Fang et al., 2026) | annotator- or pair-dependent comparison scale | RM-side L-$\varepsilon$ | binary + context or auxiliary structure | local / conditional under restrictions |
| **Comparison-specific observation-scale heterogeneity (H-DPO)** | **comparison-specific ambiguity** | **L-$\varepsilon$** | **binary pairs only** | **locally identifiable under restrictions** |

## 3 Heteroscedastic Logit RUM and Its Direct Optimization

Under the random-utility view of Section 2, standard DPO corresponds to a homoscedastic conditional-logit comparison model. The DPO-side error-layer branch identified in Section 2.4 is obtained by relaxing only this homoscedasticity assumption while keeping the DPO utility-gap parameterization unchanged. H-DPO should be read as the minimal DPO-side instantiation of comparison-specific observation-scale heterogeneity, rather than as a generic pair-adaptive objective.

### 3.1 From Standard DPO to Heteroscedastic Direct Preference Optimization

Consider a prompt $x$ and two candidate responses $y_w$ and $y_l$, where $y_w$ is observed as preferred to $y_l$. Let $V_\pi(x, y)$ denote the systematic utility of a response under policy $\pi$. Under a homoscedastic conditional-logit model (Bradley & Terry, 1952; McFadden, 1974; Train, 2009), the pairwise preference probability takes the form

$$P(y_w \succ y_l \mid x) = \sigma\big(\big(V_\pi(x, y_w) - V_\pi(x, y_l)\big)/\tau\big)\,, \tag{3}$$

where $\tau > 0$ is a single global observation scale.

Standard DPO (Rafailov et al., 2023) instantiates this model by parameterizing the utility difference through policy-relative log-probability ratios against a reference policy $\pi_{\mathrm{ref}}$. Define

$$\Delta_\pi(x, y_w, y_l) := \log\big(\pi(y_w \mid x)/\pi_{\mathrm{ref}}(y_w \mid x)\big) - \log\big(\pi(y_l \mid x)/\pi_{\mathrm{ref}}(y_l \mid x)\big). \tag{4}$$

Then DPO models the observed preference probability as

$$P(y_w \succ y_l \mid x) = \sigma(\beta\, \Delta_\pi(x, y_w, y_l))\,, \tag{5}$$

with global inverse temperature $\beta > 0$, equivalently a homoscedastic conditional-logit model with $\tau = 1/\beta$. We now relax only the single-scale assumption. In random-utility terms, the utility-error difference $\varepsilon(x, y_w) - \varepsilon(x, y_l)$ is logistic with **comparison-specific observation scale** $\tau(x, y_w, y_l) > 0$. The same latent utility gap can then induce different preference sharpness across comparisons. The resulting heteroscedastic model replaces the single global scale in Eq. (3) by a comparison-specific one, $P(y_w \succ y_l \mid x) = \sigma\big(\big(V_\pi(x, y_w) - V_\pi(x, y_l)\big)/\tau(x, y_w, y_l)\big)$, leaving the utility-gap structure otherwise unchanged. Without structural restrictions, binary preference data identify only the ratio $(V_\pi(x, y_w) - V_\pi(x, y_l))/\tau(x, y_w, y_l)$; formal identifiability is deferred to Section 4.

Retaining the DPO utility-gap parameterization and relaxing only the homoscedastic observation-scale assumption yields

$$P(y_w \succ y_l \mid x) = \sigma(\beta\, \Delta_\pi(x, y_w, y_l)/\tau(x, y_w, y_l))\,. \tag{6}$$

Given a preference dataset $\mathcal{D} = \{(x_i, y_w^i, y_l^i)\}_{i=1}^N$, the corresponding negative log-likelihood is $\mathcal{L}_{\text{H-DPO}}(\pi, \tau) = -\sum_{i=1}^N \log \sigma\big(\big(\beta\, \Delta_\pi(x_i, y_w^i, y_l^i)\big)/\tau(x_i, y_w^i, y_l^i)\big)$.

**Proposition 1** (DPO as the homoscedastic special case)**.** *If $\tau(x, y_w, y_l) \equiv c$ is constant almost surely for some $c > 0$, then H-DPO reduces to standard DPO up to a rescaling of the global inverse-temperature parameter: defining $\beta' = \beta/c$ gives $\sigma\big(\big(\beta\, \Delta_\pi(x, y_w, y_l)\big)/\tau(x, y_w, y_l)\big) = \sigma(\beta'\, \Delta_\pi(x, y_w, y_l))$.*

### 3.2 Why H-DPO Occupies the Error Layer Rather Than the Loss Layer

The key distinction is simple: a pair-dependent scalar plays a different role depending on whether it changes the latent observation channel itself or only reweights the optimization surrogate. Consider a representative loss-adaptive objective

$$\mathcal{L}_{\text{wDPO}} = -\sum_{i=1}^N w_i \log \sigma(\beta\, \Delta_i), \tag{7}$$

where $w_i > 0$ is an externally assigned pair weight and $\Delta_i := \Delta_\pi(x_i, y_w^i, y_l^i)$. This leaves the conditional preference model unchanged: all comparisons are still generated by the same homoscedastic logistic channel, and only their contribution to the empirical loss is rescaled.

H-DPO instead introduces the pair-dependent scale inside the logit:

$$\ell_i^{\text{H-DPO}} = -\log \sigma(\beta\, \Delta_i/\tau_i)\,, \qquad \tau_i := \tau(x_i, y_w^i, y_l^i). \tag{8}$$

Two comparisons with the same score $\Delta_i$ but different $\tau_i$ therefore induce different preference probabilities, not merely different weights. The distinction is also visible in the score gradient:

$$\frac{\partial \ell_i^{\text{H-DPO}}}{\partial \Delta_i} = -\frac{\beta}{\tau_i}\sigma(-\beta\Delta_i/\tau_i)\,, \qquad \frac{\partial \ell_i^{\text{wDPO}}}{\partial \Delta_i} = -w_i\beta\, \sigma(-\beta\Delta_i). \tag{9}$$

Thus H-DPO changes the conditional preference model itself, whereas weighted DPO rescales only the empirical loss contribution. This difference is structural rather than cosmetic.

**Proposition 2** (Non-equivalence to score-independent loss reweighting)**.** *Unless $\tau(x, y_w, y_l)$ is almost surely constant, the H-DPO per-comparison loss $-\log \sigma(\beta\, \Delta/\tau)$ cannot be represented for all $\Delta$ by any score-independent positive reweighting of the homoscedastic DPO loss $w(x, y_w, y_l)\,[-\log \sigma(\beta\, \Delta)]$. Consequently, the two objectives induce different estimating equations and, in general, different population stationary points; the proof is given in Appendix B.6.*

### 3.3 Practical Parameterization of the Comparison Scale

The theory in this paper concerns the heteroscedastic-logit branch as a model class; the parameterization below is the practical instantiation used in experiments. We model comparison-specific observation scale with a lightweight symmetric ambiguity head on the log scale, $\log \tau_\phi(x, y_w, y_l) = b_\phi + s_\phi(x, y_w, y_l)$, i.e. $\tau_\phi(x, y_w, y_l) = \exp\big(b_\phi + s_\phi(x, y_w, y_l)\big)$, where $b_\phi \in \mathbb{R}$ is a learnable global log-scale and $s_\phi(\cdot)$ is a comparison-specific predictor. Since only the ratio $\beta\, \Delta/\tau$ is identified, we

fix $\beta$ to its standard DPO value and let the scale head absorb the global normalization. Implementation details are deferred to Appendix E.

With this parameterization, the training objective is

$$\mathcal{L}_{\text{train}}(\pi, \phi) = -\frac{1}{N}\sum_{i=1}^{N} \log \sigma\left(\beta \, \frac{\Delta_\pi(x_i, y_w^i, y_l^i)}{\tau_i}\right) + \lambda_\tau \frac{1}{N}\sum_{i=1}^{N} \tilde{s}_\phi(z_i)^2, \tag{10}$$

where $\tau_i = \tau_\phi(x_i, y_w^i, y_l^i)$, $z_i = (x_i, y_w^i, y_l^i)$, $\tilde{s}_\phi$ is the minibatch-centered counterpart of the local log-scale $s_\phi$, and $\lambda_\tau > 0$ is a small regularization coefficient. This regularizer is a practical stabilization device rather than part of the branch-level theory.

## 4 Statistical Consequences of the Error-Layer Branch

This section formalizes the branch-level theory for the DPO-side error-layer branch highlighted in Figure 1. The central question is simple: if the true preference channel has comparison-specific observation scale, what quantity does a homoscedastic DPO score recover? Our claims concern restricted heteroscedastic-logit model classes rather than arbitrary neural parameterizations.

### 4.1 Setup and Semantic Misspecification

Let $Z = (x, y_w, y_l)$ denote a random comparison triple from distribution $\mathcal{P}$. Suppose the true preference-generation process belongs to the error-layer branch instantiated in Section 3:

$$p^\star(Z) = P^\star(y_w \succ y_l \mid Z) = \sigma\big(\big(\beta\,\Delta^\star(Z)\big)/\tau^\star(Z)\big)\,, \tag{11}$$

where $\Delta^\star(Z)$ is the branch-consistent latent utility gap of interest and $\tau^\star(Z) > 0$ is the true comparison-specific observation scale.

For any constant $c > 0$, the pairs $(\Delta^\star, \tau^\star)$ and $(c\Delta^\star, c\tau^\star)$ induce the same conditional preference probabilities. The decomposition is therefore identifiable only up to joint global scaling, and $\beta$ is treated as fixed by the global normalization throughout this section.

We write $p_\Delta(Z) = \sigma(\beta\,\Delta(Z))$ for the homoscedastic score model and $p_{\Delta,\tau}(Z) = \sigma(\beta\,\Delta(Z)/\tau(Z))$ for the heteroscedastic one. Here "semantic" refers specifically to whether the learned score remains proportional to the latent utility gap of interest, not merely whether it predicts the correct binary preferences. This motivates the definition below.

**Definition 1** (Utility-faithful specification)**.** A homoscedastic score model $p_\Delta(Z) = \sigma(\beta\,\Delta(Z))$ is *utility-faithfully specified* for the true error-layer process $p^\star(Z)$ if there exists a measurable score function $\Delta(Z)$ such that:

1. $\Delta(Z) = c\,\Delta^\star(Z)$ almost surely for some constant $c > 0$, and
2. $p_\Delta(Z) = p^\star(Z)$ almost surely.

The proportionality requirement rules out weaker monotone equivalence; see Appendix A.

**Theorem 1** (Semantic misspecification of the homoscedastic branch)**.** *Suppose the true conditional preference probabilities satisfy Eq. (11), where $\beta$ is part of the fixed global normalization, and assume that $\tau^\star(Z)$ is not almost surely constant on $\{Z : \Delta^\star(Z) \neq 0\}$. Then no homoscedastic score model $p_\Delta(Z) = \sigma(\beta\,\Delta(Z))$ is utility-faithfully specified for $p^\star(Z)$ in the sense of Definition 1.*

The proof is given in Appendix B. Theorem 1 is the central theoretical result of the paper. It is a semantic misspecification statement, not an expressivity impossibility result: a sufficiently flexible homoscedastic score may still fit observed preference probabilities, but generally only by losing utility-faithful semantics.

**Corollary 1** (Exact homoscedastic fitting recovers a distorted score)**.** *Under the assumptions of Theorem 1, if a measurable function $\Delta(Z)$ satisfies $\sigma(\beta\,\Delta(Z)) = p^\star(Z)$ almost surely, then $\Delta(Z) = \Delta^\star(Z)/\tau^\star(Z)$ almost surely.*

Corollary 1 gives the operational meaning of the misspecification: because $\beta$ is fixed by global normalization, a homoscedastic score can recover the correct probabilities only by encoding a scale-distorted score that combines utility magnitude and local observation scale. More generally, the same distortion characterizes the population minimizer of homoscedastic Bernoulli cross-entropy whenever the homoscedastic score class is rich enough to realize the probability-matching solution: the recovered score targets $\Delta^\star/\tau^\star$, not $\Delta^\star$ (Appendix B). Scale heterogeneity biasing homoscedastic-logit preference parameters is classical in discrete-choice econometrics (Fiebig et al., 2010; Hess & Train, 2017); our contribution is its consequence for DPO semantics, where the implicit reward is intended as a utility gap.

Moreover, positive loss-layer reweighting does not alter this homoscedastic population target: weighted homoscedastic objectives can change optimization dynamics and finite-sample behavior, but not the scale-distorted score recovered by the branch (Appendix B). Recovering utility-faithful semantics therefore requires changing the conditional preference model itself.

### 4.2 Restricted Identifiability and Population Consistency

Because utility magnitude and comparison-specific observation scale enter the likelihood only through their ratio, binary preference data do not globally identify arbitrary pairs $(\Delta, \tau)$, which are determined only up to the global scaling invariance $(\Delta, \tau) \sim (c\Delta, c\tau)$ noted above. The relevant claim is therefore *local* identifiability within a restricted smooth family, modulo this scaling: for a smooth family $(\Delta_\theta, \tau_\phi)$ with $\tau_\phi > 0$ and induced heteroscedastic logit field $\eta_{\theta,\phi} = \beta\, \Delta_\theta/\tau_\phi$, two distinct local equivalence classes cannot induce the same field $\eta_{\theta,\phi}$ near the true parameter. Under such restrictions, the error-layer branch is identifiable in principle from binary data, unlike branches such as user-type mixtures.

Under the true process in Eq. (11), the heteroscedastic-logit class $p_{\Delta,\tau}(Z) = \sigma(\beta\, \Delta(Z)/\tau(Z))$ is well specified by construction. The following theorem gives the corresponding population-level consistency result at the level of conditional preference probabilities and, under local identifiability, at the level of the latent decomposition up to global scaling.

**Theorem 2** (Population risk consistency of the heteroscedastic branch)**.** *Let $p_{\theta,\phi}(Z) = \sigma\big(\big(\beta\, \Delta_\theta(Z)\big)/\tau_\phi(Z)\big)$ be a heteroscedastic-logit model class, and suppose the true conditional preference probability satisfies $p^\star(Z) = p_{\theta^\star,\phi^\star}(Z)$ a.s. for some $(\theta^\star, \phi^\star)$. Define the population cross-entropy risk*

$$\mathcal{R}(\theta, \phi) = \mathbb{E}\Big[ - p^\star(Z) \log p_{\theta,\phi}(Z) - (1 - p^\star(Z)) \log(1 - p_{\theta,\phi}(Z))\Big]. \tag{12}$$

*Then every global minimizer $(\hat\theta, \hat\phi)$ of $\mathcal{R}$ satisfies $p_{\hat\theta,\hat\phi}(Z) = p^\star(Z)$ a.s. If, in addition, the family is locally identifiable up to global scaling at $(\theta^\star, \phi^\star)$, then every minimizer sufficiently close to $(\theta^\star, \phi^\star)$ belongs to the same local equivalence class as $(\theta^\star, \phi^\star)$ modulo global scaling.*

Theorem 2 is a probability-level consistency statement under realizability, not a finite-sample or optimization guarantee for the full neural implementation; unrestricted neural scale parameterizations remain outside the scope of the present theory.

## 5 Experiments

We evaluate the predictions of the DPO-side error-layer branch through synthetic validation, real-world benchmarks, and mechanism ablations.

We train on UltraFeedback (Cui et al., 2024), whose pairwise comparisons often involve partial instruction satisfaction, stylistic trade-offs, and closely matched responses, making it a natural setting for comparison-specific observation-scale heterogeneity. We evaluate *preference discrimination* on RewardBench v1 (Lambert et al., 2024) and v2 (Malik et al., 2026), and *open-ended generation quality* on AlpacaEval 2.0 (Dubois et al., 2024) and Arena-Hard (Li et al., 2024).

We compare H-DPO against three baseline groups aligned with the taxonomy of Section 2.3: (i) homoscedastic references, including DPO (Rafailov et al., 2023), IPO (Azar et al., 2024), and SimPO (Meng et al., 2024); (ii) neighboring loss-/process-layer (L-X) adaptations, including $\beta$-

Table 2: Synthetic sweep (mean ± std., five seeds). $\rho_\Delta$: utility-gap recovery. $G_{\text{dist}} = \rho_{\Delta/\tau} - \rho_\Delta$: positive values favor the distorted $\Delta^\star/\tau^\star$, negative the true $\Delta^\star$, and zero indicates a tie. NLL: the three methods agree within $0.001$ at every $\alpha$, so a single value (H-DPO) is shown. $\rho_{\log\tau}$: scale recovery (H-DPO only); at $\alpha = 0$ the true scale is constant, so it is undefined and reported as $0.000$. W-DPO uses the true $1/\tau^\star$ only as a loss weight. Bold = best $\rho_\Delta$ per row ($\alpha > 0$).

| | $\rho_\Delta \uparrow$ | | | $G_{\text{dist}}$ | | | | |
|---|---|---|---|---|---|---|---|---|
| $\alpha$ | **DPO** | **H-DPO** | **W-DPO** | **DPO** | **H-DPO** | **W-DPO** | **NLL↓** | $\rho_{\log\tau}$ |
| 0.00 | .989±.001 | .988±.001 | .989±.001 | .000±.000 | .000±.000 | .000±.000 | 0.604 | .000 |
| 0.25 | .966±.002 | **.981±.002** | .964±.003 | +.021±.002 | −.003±.004 | +.024±.003 | 0.598 | .800 |
| 0.50 | .908±.008 | **.972±.003** | .897±.008 | +.071±.010 | −.038±.008 | +.083±.010 | 0.586 | .919 |
| 0.75 | .840±.012 | **.954±.011** | .823±.008 | +.108±.018 | −.094±.028 | +.133±.008 | 0.569 | .939 |
| 1.00 | .776±.014 | **.946±.018** | .748±.020 | +.107±.034 | −.205±.045 | +.156±.038 | 0.553 | .949 |

DPO (Wu et al., 2024) and $\varepsilon$-DPO (Lee et al., 2025); and (iii) methods from other heterogeneity branches, including utility-layer (L-V) methods such as MixDPO (Imai et al., 2026) and EM-DPO (Chidambaram et al., 2025), and the temporal-layer (L-T) method NS-DPO (Son et al., 2025).

### 5.1 Synthetic Validation of the Error-Layer Branch

We generate binary preferences from the heteroscedastic-logit model with independent latent utility and scale blocks ($z_i^\Delta, z_i^\tau \sim \mathcal{N}(0, I)$), a linear gap $\Delta_i^\star = {w_\Delta^\star}^\top z_i^\Delta$, and a comparison-specific scale $\tau_i^\star = \exp(\alpha {w_\tau^\star}^\top z_i^\tau)$:

$$r_i \sim \text{Bernoulli}\big(\sigma\big(\big(\beta\Delta_i^\star\big)/\tau_i^\star\big)\big)\,. \tag{13}$$

This setup directly instantiates the branch analyzed in Section 4. We train standard DPO, H-DPO, and an oracle Weighted DPO (W-DPO) that uses the true $1/\tau_i^\star$ only as a loss weight while keeping the comparison channel homoscedastic. We report utility-gap recovery $\rho_\Delta = \text{Corr}(\hat{\Delta}, \Delta^\star)$, alignment with the scale-distorted score $\rho_{\Delta/\tau} = \text{Corr}(\hat{\Delta}, \Delta^\star/\tau^\star)$, the signed distortion gap $G_{\text{dist}} = \rho_{\Delta/\tau} - \rho_\Delta$, test NLL, and scale recovery $\rho_{\log\tau}$ (full definitions and protocol in Appendix F).

Table 2 directly matches the predictions of Section 4. When $\alpha = 0$, all three methods coincide: H-DPO reduces to DPO in the homoscedastic regime. As heterogeneity grows, standard DPO is pulled toward the scale-distorted score $\Delta^\star/\tau^\star$: $\rho_\Delta$ falls sharply while $G_{\text{dist}}$ remains positive, and H-DPO's recovered-quantity advantage over DPO is positive in every paired seed at every nonzero heterogeneity level (Appendix F.6). H-DPO instead preserves high $\rho_\Delta$, drives $G_{\text{dist}} \leq 0$, and recovers the latent scale ($\rho_{\log\tau} \geq 0.91$ for $\alpha \geq 0.5$). Test NLL is essentially identical across methods, so the difference is not preference fit but whether the model recovers the latent utility gap or a scale-distorted score. The oracle W-DPO control tracks DPO rather than H-DPO, showing that loss-layer reweighting does not alter the homoscedastic target.

### 5.2 Real-World Preference Optimization Benchmarks

Table 3 summarizes the full benchmark suite on three base models. On RewardBench, H-DPO is at or above standard DPO across all three bases: it improves over DPO on Llama-3.1-8B-Instruct and Gemma-2-9B-it, and is statistically indistinguishable from DPO on Mistral-7B-Instruct-v0.3. Generation-side point estimates are likewise generally favorable, though most are not individually significant. We therefore treat the paired contrast against standard DPO on preference discrimination as the primary empirical claim from the real-world suite, with generation-side differences reported descriptively rather than as a universal win over neighboring methods.

Additional analyses support the intended mechanism. On RewardBench v1, the learned scale remains positively associated with an external semantic-ambiguity proxy after controlling for reward margin, response lengths, and subset (partial $\rho = 0.298$, 95% CI $[0.265, 0.332]$). On RewardBench v2, more than half of the log-scale variance lies across response pairs under the same prompt (within-prompt share $0.557$; ICC $= 0.443$), indicating substantial pair-specific variation beyond prompt difficulty. Together, these findings are consistent with interpreting $\tau_\phi$ as reflecting comparison-specific ambiguity rather than merely prompt difficulty or reward margin.

Table 3: Real-world results for UltraFeedback-tuned models. **AE2**: AlpacaEval 2.0 length-controlled (LC) / raw win rate (WR); **AH**: Arena-Hard style-controlled (SC) / headline Elo win rate (WR); **RB**: RewardBench v1/v2 accuracy under the implicit-reward log-ratio. †SimPO is reference-free, so its RB entries are not commensurable (Appendix G). Bold = best per column.

| | **Llama-3.1-8B-Instruct** | | | | | | **Gemma-2-9B-it** | | | | | | **Mistral-7B-Instruct-v0.3** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **AE2** | | **AH** | | **RB** | | **AE2** | | **AH** | | **RB** | | **AE2** | | **AH** | | **RB** | |
| **Method** | **LC** | **WR** | **SC** | **WR** | **v1** | **v2** | **LC** | **WR** | **SC** | **WR** | **v1** | **v2** | **LC** | **WR** | **SC** | **WR** | **v1** | **v2** |
| DPO | 23.6 | 26.7 | 38.1 | 38.4 | 75.8 | 46.3 | 47.6 | 48.1 | 58.9 | 58.0 | 63.6 | 36.2 | 15.4 | 19.6 | 23.9 | 23.7 | 68.0 | 40.9 |
| SimPO† | 22.4 | 23.4 | 36.0 | 36.6 | 72.2 | 47.9 | 52.4 | 51.4 | 59.9 | 59.8 | 61.4 | 29.8 | 17.6 | 19.5 | 19.2 | 19.7 | 66.0 | 38.0 |
| IPO | 20.8 | 23.6 | 37.8 | 37.7 | 69.5 | 44.7 | 51.9 | 53.7 | 59.9 | 59.9 | 54.8 | 28.8 | 16.2 | 18.1 | 21.7 | 22.8 | 63.0 | 37.6 |
| $\beta$-DPO | 19.3 | 20.9 | 35.2 | 34.7 | 72.8 | 41.1 | 44.7 | 43.2 | 52.6 | 52.0 | 55.4 | 25.1 | 13.9 | 14.4 | 20.0 | 19.6 | 60.2 | 34.6 |
| $\varepsilon$-DPO | 20.9 | 24.0 | 39.1 | 39.0 | 75.2 | 47.2 | **54.7** | **55.5** | 56.9 | 56.5 | 62.3 | 35.0 | 12.3 | 18.9 | 21.4 | 21.1 | 68.1 | **42.1** |
| MixDPO | 18.7 | 21.5 | 35.0 | 34.6 | 76.4 | 40.2 | 41.4 | 39.4 | 54.4 | 53.2 | 60.9 | 30.5 | 10.8 | 10.1 | 17.1 | 16.6 | 65.8 | 38.8 |
| EM-DPO | 22.4 | 25.2 | 37.8 | 37.8 | 74.9 | 46.4 | 48.3 | 48.5 | 61.2 | 59.9 | 63.9 | 36.5 | 15.6 | 19.5 | 23.4 | 23.2 | **68.2** | 41.3 |
| NS-DPO | 22.0 | 24.2 | 39.2 | 39.2 | 75.7 | 46.1 | 51.0 | 50.6 | 61.6 | **60.4** | 63.9 | 36.3 | 15.6 | 19.9 | 23.3 | 23.6 | 68.1 | 41.0 |
| **H-DPO** | **24.4** | **28.0** | **40.6** | **40.5** | **76.8** | **48.1** | 48.8 | 48.3 | **62.1** | 60.3 | **65.0** | **37.5** | **17.7** | **22.5** | **24.7** | **24.8** | 68.1 | 41.0 |

Table 4: Ablations on Llama-3.1-8B-Instruct; metric abbreviations as in Table 3. Bold = best per column; marginal 95% intervals for all variants are in Table 17.

| **Variant** | **AE2 LC** | **AE2 WR** | **AH SC** | **AH WR** | **RB v1** | **RB v2** |
|---|---|---|---|---|---|---|
| DPO (reference) | 23.6 | 26.7 | 38.1 | 38.4 | 75.8 | 46.3 |
| **H-DPO (full)** | **24.4** | **28.0** | 40.6 | 40.5 | **76.8** | **48.1** |
| No division: additive head | 23.1 | 25.6 | 38.2 | 38.4 | 75.6 | 47.7 |
| No division: loss weight (W-DPO) | 21.2 | 22.5 | 38.5 | 37.9 | 71.0 | 43.5 |
| With division: no ref. cues | 22.6 | 25.3 | 37.8 | 38.0 | 75.6 | 47.9 |
| With division: no prompt feats. | 22.6 | 25.3 | **40.8** | **40.6** | 76.4 | 47.7 |
| With division: prompt feats. only | 21.1 | 24.1 | 36.8 | 37.1 | 75.7 | 48.0 |

### 5.3 Ablations for the Comparison-Scale Mechanism

We ablate H-DPO on Llama-3.1-8B-Instruct, holding the trainer, data, reference policy, decoding, $\beta$, and $\lambda_\tau$ fixed. The first group removes the inside-logit division while keeping the head: the additive-head variant is capacity-matched to full H-DPO and differs only in whether the learned scalar divides the preference logit, while the loss-weight variant uses the same signal only as a per-pair weight. The second keeps the division but zeros feature groups before the head: frozen-reference ambiguity cues ("no ref. cues"), prompt-conditioned features ("no prompt feats."), or all but the latter ("prompt feats. only"); see Appendix E for the full feature list.

Table 4 shows that the inside-logit mechanism matters on the discrimination axis. Removing the division weakens the gain and brings performance close to DPO, while using the same signal as a loss weight is substantially worse: it drops RewardBench by $5.82$ points on v1 and $4.58$ on v2 relative to full H-DPO, and is the only variant below plain DPO on both versions. By contrast, the feature-mask variants remain comparatively close to the full model on RewardBench, although keeping only the prompt-conditioned features also weakens generation metrics (Table 4). This pattern is consistent with the taxonomy in Section 2.3: the main effect comes from modeling comparison-specific observation scale inside the conditional preference model, with pair-level comparison features appearing to be the active ingredient rather than auxiliary capacity or generic pair-adaptive weighting.

## 6 Conclusion

We presented a unified random-utility view of heterogeneity in direct preference optimization, isolated a DPO-side branch of comparison-specific observation-scale heterogeneity, and instantiated it as H-DPO. Our central result is that homoscedastic DPO is semantically misspecified for this branch: it learns a scale-distorted utility signal; because observation scale is positive, this primarily affects magnitude rather than population pairwise ranking. We hope this view clarifies what DPO-like methods model and what binary preference data can support.

### AI USE STATEMENT

Generative AI tools were used to assist with literature discovery and citation checking; checking mathematical claims and proof arguments; and language polishing, including translation, readability editing, manuscript restructuring, and LaTeX formatting. Generative AI was not used to generate synthetic or benchmark data, to create benchmark labels, to clean or restructure the datasets, or to autonomously execute the reported experiments. The authors manually verified cited sources, derivations, proofs, code changes, experimental outputs, statistical calculations, and all resulting claims. All authors reviewed the AI-assisted work and take responsibility for the final text, claims, code, data usage, and artifacts.

### ETHICS STATEMENT

This work studies preference optimization for language models using existing preference and evaluation datasets. It introduces no new human-subject data collection. As with other alignment methods, improved preference optimization may inherit biases or limitations from the training preferences, reference policies, reward models, and evaluation judges. We therefore interpret the reported gains as benchmark-specific evidence rather than as a guarantee of universally aligned behavior.

### REPRODUCIBILITY STATEMENT

The paper states the statistical model, objectives, assumptions, and principal experimental protocol in the main text. Complete proofs and an identifiability example are provided in the supplementary material. The supplementary material also documents the scale-head features and training algorithm, synthetic data-generating process and commands, real-world training and evaluation configurations, full confidence intervals and paired tests, ablation definitions, and hyperparameter sweeps. These materials are intended to make the theoretical and empirical claims independently auditable; anonymized code and machine-readable experiment artifacts are provided as supplementary material.

## REFERENCES

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024. arXiv:2310.12036.

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

Haoxian Chen, Hanyang Zhao, Henry Lam, David D. Yao, and Wenpin Tang. MallowsPO: Fine-tune your LLM with preference dispersions. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2405.14953.

Keertana Chidambaram, Karthik Vinary Seetharaman, and Vasilis Syrgkanis. Direct preference optimization with unobserved preference heterogeneity: The necessity of ternary preferences. *arXiv preprint arXiv:2510.15716*, 2025.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. UltraFeedback: Boosting language models with scaled AI feedback. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2310.01377.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2404.04475.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. KTO: Model alignment as prospect theoretic optimization. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2402.01306.

Shuxing Fang, Ruijian Han, Liangyu Zhang, and Fan Zhou. Variance-aware reward modeling with anchor guidance. *arXiv preprint arXiv:2605.11865*, 2026.

Denzil G. Fiebig, Michael P. Keane, Jordan Louviere, and Nada Wasi. The generalized multinomial logit model: Accounting for scale and coefficient heterogeneity. *Marketing Science*, 29(3):393–421, 2010.

Hiroki Furuta, Kuang-Huei Lee, Shixiang Shane Gu, Yutaka Matsuo, Aleksandra Faust, Heiga Zen, and Izzeddin Gur. Geometric-averaged preference optimization for soft preference labels. *arXiv preprint arXiv:2409.06691*, 2024.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Chaoyue He, Xin Zhou, Di Wang, Hong Xu, Wei Liu, and Chunyan Miao. SP$^2$DPO: An LLM-assisted semantic per-pair DPO generalization. *arXiv preprint arXiv:2601.22385*, 2026.

Stephane Hess and Kenneth E. Train. Correlation and scale in mixed logit models. *Journal of Choice Modelling*, 23:1–8, 2017.

Ilgee Hong, Zichong Li, Alexander Bukharin, Yixiao Li, Haoming Jiang, Tianbao Yang, and Tuo Zhao. Adaptive preference scaling for reinforcement learning with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2406.02764.

Sam Houliston, Alizée Pace, Alexander Immer, and Gunnar Rätsch. Uncertainty-penalized direct preference optimization. In *NeurIPS Workshop on Fine-Tuning in Modern Machine Learning (FITML)*, 2024. arXiv:2410.20187.

Saki Imai, Pedram Heydari, Anthony Sicilia, Asteria Kaeberlein, Katherine Atwell, and Malihe Alikhani. MixDPO: Modeling preference strength for pluralistic alignment. *arXiv preprint arXiv:2601.06180*, 2026.

Kyuyoung Kim, Ah Jeong Seo, Hao Liu, Jinwoo Shin, and Kimin Lee. Margin matching preference optimization: Enhanced model alignment with granular feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024. arXiv:2410.03145.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024.

Sangkyu Lee, Janghoon Han, Hosung Song, Stanley Jungkyu Choi, Honglak Lee, and Youngjae Yu. KL penalty control via perturbation for direct preference optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2502.13177.

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

Jinsong Liu, Dongdong Ge, and Ruihao Zhu. Reward learning from preference with ties. *arXiv preprint arXiv:2410.05328*, 2024.

Pangpang Liu, Junwei Lu, and Will Wei Sun. Uncertainty quantification for large language model reward learning under heterogeneous human feedback. *arXiv preprint arXiv:2512.03208*, 2025.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A. Smith, Hannaneh Hajishirzi, and Nathan Lambert. RewardBench 2: Advancing reward model evaluation. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2506.01937.

Daniel McFadden. Conditional logit analysis of qualitative choice behavior. In Paul Zarembka (ed.), *Frontiers in Econometrics*, pp. 105–142. Academic Press, 1974.

Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2405.14734.

Motoki Omura, Yasuhiro Fujita, and Toshiki Kataoka. Entropy controllable direct preference optimization. *arXiv preprint arXiv:2411.07595*, 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. arXiv:2403.19159.

Rhitabrat Pokharel, Yufei Tao, and Ameeta Agrawal. CAPO: Confidence aware preference optimization learning for multilingual preferences. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2025*, pp. 1144–1156, 2025. arXiv:2511.07691.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.18290.

Hyung Gyu Rho. Margin adaptive DPO: Leveraging reward model for granular control in preference optimization. *arXiv preprint arXiv:2510.05342*, 2025.

Seongho Son, William Bankes, Sayak Ray Chowdhury, Brooks Paige, and Ilija Bogunovic. Right now, wrong then: Non-stationary direct preference optimization under preference drift. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2407.18676.

Wangtao Sun, Xiang Cheng, Xing Yu, Haotian Xu, Zhao Yang, Shizhu He, Jun Zhao, and Kang Liu. Probabilistic uncertain reward model. *arXiv preprint arXiv:2503.22480*, 2025.

Kenneth E. Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2nd edition, 2009.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. $\beta$-DPO: Direct preference optimization with dynamic $\beta$. *arXiv preprint arXiv:2407.08639*, 2024.

Junkang Wu, Xue Wang, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. AlphaDPO: Adaptive reward margin for direct preference optimization. *arXiv preprint arXiv:2410.10148*, 2025.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general chinese embeddings. *arXiv preprint arXiv:2309.07597*, 2024.

# A Why Semantic Faithfulness Requires More Than Monotone Agreement

This appendix clarifies the semantic criterion used in Definition 1. The issue is not merely whether a learned score preserves the ordering of latent utility gaps, but whether it preserves the latent utility-gap *object* associated with the branch itself. This distinction matters because Theorem 1 is a statement about score semantics, not merely about ranking consistency.

In the error-layer branch studied in this paper, the latent utility gap $\Delta^\star(Z)$ enters the conditional preference probability jointly with the comparison-specific observation scale $\tau^\star(Z)$ through the ratio $\Delta^\star(Z)/\tau^\star(Z)$. Binary preference probabilities therefore depend only on this ratio. As a result, criteria based only on probability matching or on monotone agreement with $\Delta^\star(Z)$ are too weak for the semantic question in Section 4: they do not distinguish recovery of the latent utility gap itself from recovery of a score that absorbs comparison-specific scale.

A weaker alternative would be *ordinal faithfulness*, requiring only that the learned score be a strictly increasing transform of the latent utility gap:

$$\Delta(Z) = g(\Delta^\star(Z)) \qquad \text{a.s.}$$

for some strictly increasing measurable function $g$. Such a criterion is appropriate if one cares only about ranking. It is not appropriate here, because the error-layer branch posits a specific latent decomposition: $\Delta^\star(Z)$ is the utility-gap quantity of interest, whereas $\tau^\star(Z)$ is a distinct comparison-scale quantity. Monotone agreement preserves order only; it does not preserve the magnitude semantics of the latent utility gap.

This distinction is essential for the misspecification result. Under the heteroscedastic branch, a homoscedastic score can exactly match the true conditional preference probabilities by encoding

$$\Delta_{\mathrm{dist}}(Z) := \Delta^\star(Z)/\tau^\star(Z),$$

as stated in Corollary 1. In general, $\Delta_{\mathrm{dist}}(Z)$ mixes utility magnitude with local observation scale and therefore does not represent the latent utility gap itself. If semantic faithfulness were defined only up to arbitrary monotone transformation, then exact probability matching by an order-compatible surrogate would not be excluded. That notion would therefore be too weak to certify recovery of the latent utility-gap object studied in this paper.

Definition 1 therefore requires proportional recovery:

$$\Delta(Z) = c\,\Delta^\star(Z) \qquad \text{a.s. for some } c > 0.$$

This matches the natural invariance of the branch. Since $(\Delta^\star, \tau^\star)$ and $(c\Delta^\star, c\tau^\star)$ induce the same conditional preference probabilities, the latent decomposition is identifiable only up to joint global scaling. Proportional recovery preserves exactly this unavoidable ambiguity and no more: it allows global renormalization, but it does not allow comparison-specific scale variation to be absorbed into the score.

The following proposition shows that exact probability matching alone does not imply recovery of the latent utility-gap object.

**Proposition 3** (Exact probability matching does not imply proportional recovery)**.** *Suppose the true conditional preference probability satisfies Eq. (11), and assume that $\tau^\star(Z)$ is not almost surely constant on $\{Z : \Delta^\star(Z) \neq 0\}$. Define*

$$\Delta_{\mathrm{dist}}(Z) := \Delta^\star(Z)/\tau^\star(Z).$$

*Then:*

1. *the homoscedastic score model $p_\Delta(Z) = \sigma(\beta\,\Delta(Z))$ with $\Delta = \Delta_{\mathrm{dist}}$ satisfies*
$$p_\Delta(Z) = p^\star(Z) \qquad \textit{a.s.},$$
*and*
2. *there is no constant $c > 0$ such that*
$$\Delta_{\mathrm{dist}}(Z) = c\,\Delta^\star(Z) \qquad \textit{a.s. on } \{Z : \Delta^\star(Z) \neq 0\}.$$

*Proof.* By Eq. (11),

$$p^\star(Z) = \sigma(\beta\,\Delta^\star(Z)/\tau^\star(Z)) = \sigma(\beta\,\Delta_{\mathrm{dist}}(Z)) \qquad \text{a.s.},$$

which proves the first claim.

For the second claim, suppose there exists $c > 0$ such that

$$\Delta^\star(Z)/\tau^\star(Z) = c\,\Delta^\star(Z) \qquad \text{a.s. on } \{Z : \Delta^\star(Z) \neq 0\}.$$

On $\{Z : \Delta^\star(Z) \neq 0\}$, division by $\Delta^\star(Z)$ is valid, yielding

$$1/\tau^\star(Z) = c \qquad \text{a.s.},$$

so $\tau^\star(Z)$ is almost surely constant on $\{Z : \Delta^\star(Z) \neq 0\}$, contradicting the assumption. □

Proposition 3 clarifies why Definition 1 uses proportional recovery rather than a weaker criterion based only on ranking or exact probability fit. For the error-layer branch, the relevant notion of semantic faithfulness must preserve the utility-gap quantity up to the same global scaling ambiguity under which the branch itself is identifiable. In this sense, Theorem 1 does not rely on an arbitrary strengthening of faithfulness; it uses the weakest notion that still preserves the branch-level utility-gap semantics.

# B Proofs for the Main Theoretical Results

This appendix collects proofs of the main theoretical results stated in Section 4. As in the main text, these are branch-level statements for restricted heteroscedastic-logit families rather than claims about arbitrary neural parameterizations.

## B.1 Proof of Theorem 1

*Proof of Theorem 1.* Assume for contradiction that there exists a measurable score function $\Delta(Z)$ such that the homoscedastic score model $p_\Delta(Z) = \sigma(\beta\,\Delta(Z))$ is utility-faithfully specified for $p^\star(Z)$ in the sense of Definition 1. Then, by definition, there exists a constant $c > 0$ such that

$$\Delta(Z) = c\,\Delta^\star(Z) \qquad \text{a.s.}, \tag{14}$$

and

$$p_\Delta(Z) = p^\star(Z) \qquad \text{a.s.} \tag{15}$$

Using Eq. (11) and the definition of $p_\Delta$, Eq. (15) implies

$$\sigma(\beta\,\Delta(Z)) = \sigma(\beta\,\Delta^\star(Z)/\tau^\star(Z)) \qquad \text{a.s.} \tag{16}$$

Since the logistic link $\sigma$ is strictly increasing and hence injective, it follows that

$$\beta\,\Delta(Z) = \beta\,\Delta^\star(Z)/\tau^\star(Z) \qquad \text{a.s.} \tag{17}$$

Because $\beta > 0$, we obtain

$$\Delta(Z) = \Delta^\star(Z)/\tau^\star(Z) \qquad \text{a.s.} \tag{18}$$

Combining Eqs. (14) and (18) yields

$$c\,\Delta^\star(Z) = \Delta^\star(Z)/\tau^\star(Z) \qquad \text{a.s.} \tag{19}$$

Hence, on the set $\{Z : \Delta^\star(Z) \neq 0\}$,

$$c = 1/\tau^\star(Z) \qquad \text{a.s.} \tag{20}$$

Therefore $\tau^\star(Z)$ is almost surely constant on $\{Z : \Delta^\star(Z) \neq 0\}$, contradicting the hypothesis. This contradiction proves that no homoscedastic score model can be utility-faithfully specified for $p^\star(Z)$. □

## B.2 Proof of Corollary 1

*Proof of Corollary 1.* Assume that a measurable function $\Delta(Z)$ satisfies

$$\sigma(\beta\,\Delta(Z)) = p^\star(Z) \qquad \text{a.s.} \tag{21}$$

Using Eq. (11), Eq. (21) becomes

$$\sigma(\beta\,\Delta(Z)) = \sigma(\beta\,\Delta^\star(Z)/\tau^\star(Z)) \qquad \text{a.s.} \tag{22}$$

Since $\sigma$ is strictly increasing and hence injective, and $\beta > 0$, it follows that

$$\Delta(Z) = \Delta^\star(Z)/\tau^\star(Z) \qquad \text{a.s.} \tag{23}$$

Thus any exact homoscedastic fit must recover the distorted score $\Delta^\star(Z)/\tau^\star(Z)$. □

## B.3 Proof of Theorem 2

*Proof of Theorem 2.* We first prove the probability-level claim and then the local parameter-level claim.

Recall the population cross-entropy risk

$$\mathcal{R}(\theta, \phi) = \mathbb{E}\Big[ - p^\star(Z) \log p_{\theta,\phi}(Z) - (1 - p^\star(Z)) \log(1 - p_{\theta,\phi}(Z)) \Big]. \tag{24}$$

Let $q(Z) := p_{\theta,\phi}(Z)$. Since $q(Z) = \sigma(\eta_{\theta,\phi}(Z)) \in (0,1)$ a.s., the pointwise cross-entropy and KL terms below are well defined. By assumption, $p^\star(Z) = p_{\theta^\star,\phi^\star}(Z)$ a.s., so the model is realizable.

For each $Z$, define the pointwise Bernoulli cross-entropy

$$\ell_Z(q; p^\star) := -p^\star(Z) \log q(Z) - (1 - p^\star(Z)) \log(1 - q(Z)). \tag{25}$$

Then

$$\mathcal{R}(\theta, \phi) = \mathbb{E}\big[\ell_Z(q; p^\star)\big]. \tag{26}$$

Using the standard decomposition of cross-entropy into entropy plus Kullback–Leibler divergence,

$$\begin{aligned}\ell_Z(q; p^\star) = &-p^\star(Z) \log p^\star(Z) - (1 - p^\star(Z)) \log(1 - p^\star(Z)) \\ &+ \mathrm{KL}(\mathrm{Bern}(p^\star(Z)) \,\|\, \mathrm{Bern}(q(Z)))\,,\end{aligned} \tag{27}$$

so

$$\mathcal{R}(\theta, \phi) = \mathcal{R}^\star + \mathbb{E}[\mathrm{KL}(\mathrm{Bern}(p^\star(Z)) \,\|\, \mathrm{Bern}(p_{\theta,\phi}(Z)))]\,, \tag{28}$$

where

$$\mathcal{R}^\star := \mathbb{E}\Big[ - p^\star(Z) \log p^\star(Z) - (1 - p^\star(Z)) \log(1 - p^\star(Z))\Big] \tag{29}$$

does not depend on $(\theta, \phi)$.

Because the Bernoulli KL divergence is nonnegative and vanishes if and only if its two arguments are equal, Eq. (28) implies that

$$\mathcal{R}(\theta, \phi) \geq \mathcal{R}^\star \tag{30}$$

for all $(\theta, \phi)$, with equality if and only if

$$p_{\theta,\phi}(Z) = p^\star(Z) \qquad \text{a.s.} \tag{31}$$

Since $p^\star(Z) = p_{\theta^\star,\phi^\star}(Z)$ a.s. by assumption, $(\theta^\star, \phi^\star)$ attains the value $\mathcal{R}^\star$. Hence global minimizers exist, and every global minimizer $(\hat{\theta}, \hat{\phi})$ of $\mathcal{R}$ satisfies

$$p_{\hat{\theta},\hat{\phi}}(Z) = p^\star(Z) \qquad \text{a.s.} \tag{32}$$

This proves the first claim.

For the second claim, assume in addition that the family is locally identifiable up to global scaling at $(\theta^\star, \phi^\star)$. Let $(\hat{\theta}, \hat{\phi})$ be any global minimizer sufficiently close to $(\theta^\star, \phi^\star)$. By the first part,

$$p_{\hat{\theta},\hat{\phi}}(Z) = p^\star(Z) = p_{\theta^\star,\phi^\star}(Z) \qquad \text{a.s.} \tag{33}$$

Since the logistic link is injective, this is equivalent to equality of the induced heteroscedastic logit fields almost surely:

$$\eta_{\hat{\theta},\hat{\phi}}(Z) = \eta_{\theta^\star,\phi^\star}(Z) \qquad \text{a.s.} \tag{34}$$

By the local-identifiability-up-to-scaling assumption stated in Section 4.2, equality of the induced fields in a neighborhood of $(\theta^\star, \phi^\star)$ implies that $(\hat{\theta}, \hat{\phi})$ belongs to the same local equivalence class as $(\theta^\star, \phi^\star)$ modulo global scaling. □

### B.4 Estimation Distortion of the Homoscedastic Fit

The following corollary makes precise the sense in which an exact homoscedastic fit to the heteroscedastic branch recovers the distorted target $\Delta^\star/\tau^\star$ rather than the latent utility gap $\Delta^\star$, both at the population level and for the empirical risk minimizer; it formalizes the population-minimizer claim stated in the main text immediately after Corollary 1. Throughout this subsection and the next, let $\mathfrak{D}$ denote a class of measurable scores that contains $\Delta^\star/\tau^\star$.

**Corollary 2** (Estimation distortion of the homoscedastic fit)**.** *Let $Y \mid Z \sim \mathrm{Bernoulli}(p^\star(Z))$ and $\mathcal{R}_{\mathrm{homo}}(\Delta) = \mathbb{E}[\ell_{\mathrm{CE}}(\Delta; Z)]$, where $\ell_{\mathrm{CE}}$ is the pointwise Bernoulli cross-entropy. If $\Delta^\star/\tau^\star \in \mathfrak{D}$, then $\mathcal{R}_{\mathrm{homo}}$ is minimized over all measurable scores uniquely at $\Delta^\star/\tau^\star$. If in addition $\mathcal{R}_{\mathrm{homo}}$ has a well-separated minimum on $\mathfrak{D}$ and $\{\ell_{\mathrm{CE}}(\Delta; \cdot) : \Delta \in \mathfrak{D}\}$ obeys a uniform law of large numbers with an integrable envelope, then any empirical minimizer $\hat{\Delta}_n \in \arg\min_{\mathfrak{D}} \frac{1}{n} \sum_i \ell_{\mathrm{CE}}(\Delta; Z_i)$ satisfies $\mathcal{R}_{\mathrm{homo}}(\hat{\Delta}_n) \to_p \min_{\mathfrak{D}} \mathcal{R}_{\mathrm{homo}}$ and $\sigma(\beta \hat{\Delta}_n) \to_p p^\star$ in $L^1(\mathcal{P})$.*

*Proof. Population uniqueness.* Let $Y = \mathbf{1}\{y_w \succ y_l\}$ with $Y \mid Z \sim \mathrm{Bernoulli}(p^\star(Z))$, so that $\mathcal{R}_{\mathrm{homo}}(\Delta) = \mathbb{E}[\ell_{\mathrm{CE}}(\Delta; Z)]$ is the marginal expectation of the pointwise cross-entropy loss. For any measurable score $\Delta$,

$$\mathcal{R}_{\mathrm{homo}}(\Delta) - \mathcal{R}_{\mathrm{homo}}(\Delta^\star/\tau^\star) = \mathbb{E}\Big[\mathrm{KL}\big(\mathrm{Bern}(p^\star(Z)) \,\big\|\, \mathrm{Bern}(\sigma(\beta\Delta(Z)))\big)\Big], \tag{35}$$

because $\mathcal{R}_{\mathrm{homo}}(\Delta^\star/\tau^\star)$ equals the conditional entropy $\mathbb{E}[-p^\star \log p^\star - (1-p^\star)\log(1-p^\star)]$: indeed $\sigma(\beta\Delta^\star/\tau^\star) = p^\star$ a.s. by Eq. (11). The Bernoulli cross-entropy is a strictly proper scoring rule (Gneiting & Raftery, 2007), so the integrand is nonnegative and vanishes only when $\sigma(\beta\Delta(Z)) = p^\star(Z)$ almost surely. Since $\sigma$ is strictly increasing and $\beta > 0$, this holds if and only if $\Delta(Z) = \Delta^\star(Z)/\tau^\star(Z)$ almost surely. Hence $\mathcal{R}_{\mathrm{homo}}$ is minimized over all measurable scores uniquely at $\Delta^\star/\tau^\star$, which lies in $\mathfrak{D}$ by assumption.

*Empirical consistency.* Assume $\mathcal{R}_{\mathrm{homo}}$ has a well-separated minimum on $\mathfrak{D}$ and $\{\ell_{\mathrm{CE}}(\Delta;\cdot) : \Delta \in \mathfrak{D}\}$ obeys a uniform law of large numbers with an integrable envelope, so that $\sup_{\Delta\in\mathfrak{D}} \big|\frac{1}{n}\sum_{i=1}^n \ell_{\mathrm{CE}}(\Delta; Z_i) - \mathcal{R}_{\mathrm{homo}}(\Delta)\big| \to_p 0$. By the standard consistency theorem for $M$-estimators under a uniformly converging criterion and a well-separated population minimum (van der Vaart, 1998, Ch. 5), any empirical minimizer $\hat{\Delta}_n \in \arg\min_{\mathfrak{D}} \frac{1}{n}\sum_i \ell_{\mathrm{CE}}(\Delta; Z_i)$ satisfies $\mathcal{R}_{\mathrm{homo}}(\hat{\Delta}_n) \to_p \min_{\mathfrak{D}} \mathcal{R}_{\mathrm{homo}} = \mathcal{R}_{\mathrm{homo}}(\Delta^\star/\tau^\star)$.

*$L^1$ convergence of the induced probabilities.* The risk gap equals the expected KL divergence, so $\mathbb{E}[\mathrm{KL}(\mathrm{Bern}(p^\star) \,\|\, \mathrm{Bern}(\sigma(\beta\hat{\Delta}_n)))] \to_p 0$. For Bernoulli laws the total variation distance is $|p-q|$, and Pinsker's inequality gives $|p-q| \le \sqrt{\mathrm{KL}/2}$ pointwise. Therefore, by Jensen,

$$\mathbb{E}\big|\sigma(\beta\hat{\Delta}_n(Z)) - p^\star(Z)\big| \le \sqrt{\tfrac{1}{2}\,\mathbb{E}\big[\mathrm{KL}(\mathrm{Bern}(p^\star) \,\|\, \mathrm{Bern}(\sigma(\beta\hat{\Delta}_n)))\big]} \to_p 0, \tag{36}$$

that is, $\sigma(\beta\hat{\Delta}_n) \to_p p^\star$ in $L^1(\mathcal{P})$. We claim only convergence in risk and in $L^1$ of the induced probabilities; uniform convergence of $\hat{\Delta}_n$ to $\Delta^\star/\tau^\star$ as functions would additionally require entropy or bracketing conditions on $\mathfrak{D}$, which we do not impose. □

### B.5 Weighting Invariance of the Population Target

**Proposition 4** (Weighting invariance of the population target)**.** *For any measurable weight $w(Z) > 0$, let $\mathcal{R}_w(\Delta) = \mathbb{E}[w(Z)\,\ell_{\mathrm{CE}}(\Delta; Z)]$. If $\sigma(\beta\Delta^\star/\tau^\star) = p^\star$ almost surely and $\Delta^\star/\tau^\star \in \mathfrak{D}$, then for every measurable score $\Delta$, $\mathcal{R}_w(\Delta) - \mathcal{R}_w(\Delta^\star/\tau^\star) = \mathbb{E}\big[w(Z)\,\mathrm{KL}\big(\mathrm{Bern}(p^\star(Z)) \,\big\|\, \mathrm{Bern}(\sigma(\beta\Delta(Z)))\big)\big] \ge 0$, with equality if and only if $\sigma(\beta\Delta) = p^\star$ almost surely. Hence the population minimizer of every positively reweighted homoscedastic objective over the rich class is $\Delta^\star/\tau^\star$, independent of $w$.*

*Proof.* Let $Y \mid Z \sim \mathrm{Bernoulli}(p^\star(Z))$ and write the pointwise cross-entropy as $\ell_{\mathrm{CE}}(\Delta; Z) = -p^\star(Z)\log\sigma(\beta\Delta(Z)) - (1-p^\star(Z))\log(1-\sigma(\beta\Delta(Z)))$, so that $\mathcal{R}_w(\Delta) = \mathbb{E}[w(Z)\,\ell_{\mathrm{CE}}(\Delta; Z)]$. Since $\sigma(\beta\Delta^\star/\tau^\star) = p^\star$ almost surely by Eq. (11), evaluating at $\Delta^\star/\tau^\star \in \mathfrak{D}$ gives, for any measurable $\Delta$,

$$\mathcal{R}_w(\Delta) - \mathcal{R}_w(\Delta^\star/\tau^\star) = \mathbb{E}\Big[w(Z)\,\mathrm{KL}\big(\mathrm{Bern}(p^\star(Z)) \,\big\|\, \mathrm{Bern}(\sigma(\beta\Delta(Z)))\big)\Big]. \tag{37}$$

Because $w(Z) > 0$ almost surely and the Bernoulli cross-entropy is a strictly proper scoring rule (Gneiting & Raftery, 2007), the integrand is almost surely nonnegative and vanishes only where $\sigma(\beta\Delta(Z)) = p^\star(Z)$. Hence $\mathcal{R}_w(\Delta) \ge \mathcal{R}_w(\Delta^\star/\tau^\star)$ with equality if and only if $\sigma(\beta\Delta(Z)) = p^\star(Z)$ almost surely. As $\sigma$ is strictly increasing and $\beta > 0$, this holds if and only if $\Delta(Z) = \Delta^\star(Z)/\tau^\star(Z)$ almost surely. Every minimizer of $\mathcal{R}_w$ over $\mathfrak{D}$ therefore coincides with $\Delta^\star/\tau^\star$ almost surely, independently of the particular positive weight $w$. Since $\tau^\star$ is not almost surely constant on $\{\Delta^\star \neq 0\}$, this minimizer is not proportional to $\Delta^\star$ there, so it fails condition (1) of Definition 1 and is not utility-faithful. Weighting can rescale the contribution of individual comparisons to the empirical average, but it cannot move the population minimizer off $\Delta^\star/\tau^\star$; only replacing the homoscedastic conditional model $\sigma(\beta\Delta)$ by the heteroscedastic model $\sigma(\beta\Delta/\tau)$, which changes the pointwise minimizer itself, alters the population target. □

### B.6 Proof of Proposition 2

*Proof of Proposition 2.* We prove two related claims. First, for a fixed comparison, the H-DPO loss profile

$$-\log \sigma(\beta\Delta/\tau)$$

cannot be represented for all scores $\Delta$ by any score-independent positive reweighting of the homoscedastic DPO loss, except in the trivial case $\tau = 1$. Second, a nontrivially reweighted homoscedastic loss is not itself the negative log-likelihood of any one-dimensional logistic preference model in the score. The first claim is the one used in Proposition 2; the second is a further structural clarification.

*Part I: profile equality with score-independent reweighting.* Fix a comparison and suppress its dependence in $\tau$. Consider

$$f(\Delta) := -\log \sigma(\beta\Delta/\tau) = \mathrm{softplus}(-\beta\Delta/\tau), \tag{38}$$

$$g(\Delta) := w\,[-\log \sigma(\beta\Delta)] + c = w\ \mathrm{softplus}(-\beta\Delta) + c, \tag{39}$$

where $w > 0$ is score-independent and $c \in \mathbb{R}$ is an arbitrary additive constant. Suppose $f(\Delta) = g(\Delta)$ for all $\Delta \in \mathbb{R}$. Then their derivatives agree everywhere. Since $\mathrm{softplus}'(t) = \sigma(t)$,

$$f'(\Delta) = -\frac{\beta}{\tau}\sigma(-\beta\Delta/\tau), \qquad g'(\Delta) = -w\beta\,\sigma(-\beta\Delta). \tag{40}$$

Equating derivatives and cancelling $\beta > 0$ gives

$$w = \frac{\sigma(-\beta\Delta/\tau)}{\tau\,\sigma(-\beta\Delta)} \qquad \text{for all } \Delta \in \mathbb{R}. \tag{41}$$

The right-hand side must therefore be constant in $\Delta$. Taking $\Delta \to +\infty$ and using $\sigma(-t) \sim e^{-t}$ yields

$$w = \lim_{\Delta\to+\infty} \frac{\sigma(-\beta\Delta/\tau)}{\tau\,\sigma(-\beta\Delta)} = \lim_{\Delta\to+\infty} \frac{1}{\tau} e^{\beta\Delta(1-1/\tau)}. \tag{42}$$

If $\tau > 1$, the limit is $+\infty$; if $\tau < 1$, the limit is $0$. Hence a finite positive constant $w$ is possible only when $\tau = 1$. In that case Eq. (41) gives $w = 1$, and then $f \equiv g$ forces $c = 0$. Conversely, when $\tau = 1$, the choice $w = 1, c = 0$ indeed gives $f \equiv g$.

Thus, for any comparison with $\tau \neq 1$, there is no score-independent $w > 0$ and no additive constant $c$ such that

$$-\log \sigma(\beta\Delta/\tau) = w[-\log \sigma(\beta\Delta)] + c \qquad \text{for all } \Delta.$$

Therefore, if $\tau(x, y_w, y_l)$ is not almost surely constant, the H-DPO per-comparison loss cannot be represented almost surely by score-independent positive reweighting of the homoscedastic DPO loss. This does not contradict Proposition 1, which allows a rescaling of the global inverse-temperature parameter $\beta$; here $\beta$ is held fixed, and the question is whether the same loss profile can be reproduced solely by score-independent reweighting.

*Part II: a nontrivially reweighted homoscedastic loss is not a logistic likelihood.* We now show that for $w \neq 1$, the profile $w[-\log \sigma(\beta\Delta)]$ is not the negative log-likelihood of any one-dimensional logistic preference model in the score, i.e. of any model $\sigma(a\Delta + d)$ with $a > 0$ and $d \in \mathbb{R}$. Suppose for contradiction that

$$-w \log \sigma(\beta\Delta) = -\log \sigma(a\Delta + d) \qquad \text{for all } \Delta \in \mathbb{R}. \tag{43}$$

Equivalently,

$$\sigma(\beta\Delta)^w = \sigma(a\Delta + d) \qquad \text{for all } \Delta \in \mathbb{R}. \tag{44}$$

Taking logarithms and sending $\Delta \to -\infty$, where $\log \sigma(t) \sim t$, yields $w\beta\Delta \sim a\Delta + d$, hence $a = w\beta$ and $d = 0$. Sending $\Delta \to +\infty$, where $\log \sigma(t) \sim -e^{-t}$, gives asymptotically

$$-w\,e^{-\beta\Delta} \quad \text{and} \quad -e^{-w\beta\Delta}, \tag{45}$$

which can agree for all sufficiently large $\Delta$ only if $w = 1$. Therefore no such $(a, d)$ exists when $w \neq 1$. Hence a nontrivially reweighted homoscedastic loss is not the negative log-likelihood of any one-dimensional logistic preference model in the score. □

# C What "Locally Identifiable up to Global Scaling" Means

This appendix clarifies the identifiability claim used in Section 4.2. Its purpose is interpretive rather than ambitious: we do not assert exact identifiability of the full neural H-DPO parameterization used in practice. Rather, the claim is a *branch-level* one: within a restricted smooth heteroscedastic-logit family, binary preference data can locally identify the latent score–scale decomposition up to the unavoidable global scaling symmetry of the branch.

## C.1 Why a scaling quotient is necessary

Under the heteroscedastic-logit branch, the conditional preference probability takes the form

$$p_{\Delta,\tau}(Z) = \sigma(\beta\,\Delta(Z)/\tau(Z))\,, \qquad \tau(Z) > 0. \tag{46}$$

For any constant $c > 0$, the pair $(\Delta, \tau)$ and the rescaled pair $(c\Delta, c\tau)$ induce the same conditional preference probabilities:

$$\sigma(\beta\,\Delta(Z)/\tau(Z)) = \sigma(\beta\,c\Delta(Z)/(c\tau(Z)))\,. \tag{47}$$

Thus the decomposition is never identifiable in an absolute sense. The natural object of identification is therefore not an individual pair $(\Delta, \tau)$, but its equivalence class under the global scaling relation

$$(\Delta, \tau) \sim (c\Delta, c\tau), \qquad c > 0. \tag{48}$$

This is the same global ambiguity already noted in Section 4.1. The phrase "up to global scaling" means precisely that identifiability is understood only modulo the equivalence relation in Eq. (48).

## C.2 Parametric setup and field-level representation

Let $\Delta_\theta(Z)$ and $\tau_\phi(Z) > 0$ be smooth parametric families indexed by $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ and $\phi \in \Phi \subset \mathbb{R}^{d_\phi}$. The induced heteroscedastic-logit model is

$$p_{\theta,\phi}(Z) = \sigma(\eta_{\theta,\phi}(Z))\,, \qquad \eta_{\theta,\phi}(Z) := \beta\,\Delta_\theta(Z)/\tau_\phi(Z). \tag{49}$$

Because the Bernoulli likelihood depends on $(\theta, \phi)$ only through the logit field $\eta_{\theta,\phi}$, the identifiability question naturally separates into two parts:

1. **structural identifiability:** whether two nearby parameter settings can induce the same heteroscedastic-logit field except for the trivial global scaling symmetry; and
2. **local statistical visibility:** whether those quotient directions remain non-degenerate under the Bernoulli observation model.

This separation is useful because the first concerns uniqueness of the latent decomposition modulo scaling, whereas the second concerns whether those locally distinct quotient directions remain visible to likelihood-based inference. The broader branch-level interpretation in the main text appeals to both.

## C.3 Local identifiability modulo global scaling

We now state the structural local notion used in the main text.

**Definition 2** (Local identifiability up to global scaling)**.** Fix $(\theta^\star, \phi^\star)$. A smooth heteroscedastic-logit family

$$\big\{(\Delta_\theta, \tau_\phi) : \theta \in \Theta,\ \phi \in \Phi\big\}$$

is *locally identifiable up to global scaling* at $(\theta^\star, \phi^\star)$ if there exists a neighborhood $U$ of $(\theta^\star, \phi^\star)$ such that whenever $(\theta, \phi) \in U$ satisfies

$$\eta_{\theta,\phi}(Z) = \eta_{\theta^\star,\phi^\star}(Z) \qquad \text{a.s.}, \tag{50}$$

the induced function pairs $(\Delta_\theta, \tau_\phi)$ and $(\Delta_{\theta^\star}, \tau_{\phi^\star})$ belong to the same equivalence class under the global scaling relation in Eq. (48). Equivalently, there exists $c > 0$ such that

$$\Delta_\theta(Z) = c\,\Delta_{\theta^\star}(Z), \qquad \tau_\phi(Z) = c\,\tau_{\phi^\star}(Z) \qquad \text{a.s.}$$

Definition 2 is a local quotient notion. It does *not* require global uniqueness over the entire parameter space, and it does *not* claim that the raw parameters $(\theta, \phi)$ themselves are unique. It requires only that, near the truth, equality of the induced heteroscedastic-logit field forces equality of the latent decomposition modulo the unavoidable scaling symmetry.

### C.4 A SUFFICIENT TEMPLATE

The following template makes precise the branch-level sufficient conditions invoked in Section 4.2. Consider a smooth heteroscedastic-logit family $(\Delta_\theta, \tau_\phi)$ with $\tau_\phi(Z) > 0$ a.s., and let $\eta_{\theta,\phi}$ be defined as in Eq. (49). Suppose that, at $(\theta^\star, \phi^\star)$,

1. after quotienting by the global scaling relation in Eq. (48), the map $(\theta, \phi) \mapsto \eta_{\theta,\phi}$ is locally injective; and
2. in a local chart of the quotient parameterization, the Bernoulli model induced by $p_{\theta,\phi}(Z) = \sigma(\eta_{\theta,\phi}(Z))$ has non-singular Fisher information.

Then condition 1 implies local identifiability up to global scaling in the sense of Definition 2, while condition 2 ensures that the corresponding quotient directions are locally non-degenerate for likelihood-based inference.

The first condition is the structural part of the quotient argument: if two nearby parameter settings induce the same field $\eta_{\theta,\phi}$, then they must lie in the same quotient class. The second condition is a local statistical regularity condition. After fixing a local chart on the quotient, non-singular Fisher information rules out locally flat observational directions beyond the scaling symmetry already quotiented out.

This sufficient template should be read as a branch-level clarification, not as a theorem about arbitrary neural parameterizations. Its role is to formalize what the main text means by saying that the heteroscedastic branch is locally identifiable from binary preference data under structural restrictions.

### C.5 WHAT IS AND IS NOT BEING CLAIMED

To avoid over-interpretation, we summarize the scope of the identifiability claim.

**What is being claimed.** Within a restricted smooth heteroscedastic-logit family, if the score–scale parameterization is locally injective modulo the global scaling symmetry and the induced Bernoulli model is locally non-degenerate, then binary preference data support local recovery of the latent decomposition up to global scaling as a branch-level identifiability statement.

**What is not being claimed.** We do not claim:

1. global identifiability over arbitrary score/scale function classes;
2. exact identifiability of the full neural H-DPO parameterization used in experiments;
3. identifiability without structural restrictions; or
4. identifiability beyond the global scaling ambiguity inherent to the branch.

These distinctions are important. Without structural restrictions, binary preference data identify only the ratio $\Delta(Z)/\tau(Z)$, not its numerator and denominator separately. The point of the branch-level claim is that this confounding need not be fatal in principle: under suitable restrictions, the decomposition can be locally recovered modulo the one symmetry that cannot be removed.

### C.6 RELATION TO THE WORKED EXAMPLE

The sufficient template above is intentionally abstract. The next appendix instantiates it in a simple finite-dimensional family to show that the assumptions are not vacuous. The worked example is illustrative rather than exhaustive: its role is to make the quotient-by-scaling argument and the local injectivity condition concrete in a highly restricted family, not to establish local identifiability for general feature-dependent scale parameterizations or to characterize the full neural model used in practice.

# D A WORKED EXAMPLE OF LOCAL IDENTIFIABILITY IN A GENUINELY HETEROSCEDASTIC FAMILY

This appendix gives a simple finite-dimensional example illustrating the identifiability discussion of Appendix C. Unlike a constant-scale sanity check, the example below is *genuinely heteroscedastic*: the comparison scale varies with $Z$. The goal remains modest. We do not attempt to characterize the full neural H-DPO parameterization used in practice, nor do we establish local identifiability for general feature-dependent scale families. Rather, we exhibit a highly restricted heteroscedastic family in which the score–scale decomposition is locally identifiable from the induced logit field. This shows that the branch-level identifiability claim in Section 4.2 is non-vacuous even when the scale is comparison-dependent.

## D.1 A MINIMAL HETEROSCEDASTIC FAMILY

Let $h(Z) \in \mathbb{R}$ be a fixed measurable scalar feature. Consider the family

$$\Delta_a(Z) = a\, h(Z), \tag{51}$$

$$\tau_b(Z) = \exp\big(b\, h(Z)\big), \tag{52}$$

where $a \in \mathbb{R} \setminus \{0\}$ and $b \in \mathbb{R}$. The induced heteroscedastic-logit field is

$$\eta_{a,b}(Z) = \beta\, \frac{\Delta_a(Z)}{\tau_b(Z)} = \beta\, a\, h(Z) \exp\big(-b\, h(Z)\big). \tag{53}$$

Whenever $b \neq 0$ and $h(Z)$ is nonconstant, the scale $\tau_b(Z)$ varies with $Z$, so this is a genuinely comparison-dependent scale family.

The branch-level global scaling symmetry acts on function pairs as

$$(\Delta, \tau) \sim (c\Delta, c\tau), \qquad c > 0.$$

In the present family, however, this symmetry is fixed by construction: multiplying $\tau_b(Z) = \exp(bh(Z))$ by a non-unit constant generally takes it outside the family. Thus the family is normalized strongly enough that exact local identifiability is possible. Since exact identifiability is stronger than quotient identifiability, it will in particular imply local identifiability up to global scaling in the sense of Definition 2.

## D.2 A REGULARITY CONDITION

We work around a reference point $(a^\star, b^\star)$ under the following mild non-degeneracy condition.

**Assumption 1** (Regularity of the worked heteroscedastic family)**.** The essential range of $h(Z)$ contains at least two distinct nonzero points.

This assumption has two roles. First, the presence of at least two distinct nonzero values of $h(Z)$ allows the scale parameter $b$ to be distinguished from the score amplitude $a$ on the set $\{h(Z) \neq 0\}$, where the identification argument operates. Second, requiring these values to be nonzero rules out the trivial zero-field case.

## D.3 STRUCTURAL IDENTIFIABILITY OF THE DECOMPOSITION

We first verify the structural part of the branch-level template.

**Lemma 1** (Local injectivity of the heteroscedastic decomposition)**.** *Under Assumption 1, suppose $(a, b)$ and $(a', b')$ are sufficiently close to $(a^\star, b^\star)$ and satisfy*

$$\eta_{a,b}(Z) = \eta_{a',b'}(Z) \qquad \text{a.s.} \tag{54}$$

*Then $a = a'$ and $b = b'$. In particular, the family is locally identifiable, hence locally identifiable up to global scaling.*

*Proof.* Using Eq. (53), Eq. (54) means

$$a\, h(Z) \exp\big(-b\, h(Z)\big) = a'\, h(Z) \exp\big(-b'\, h(Z)\big) \qquad \text{a.s.} \tag{55}$$

On the set $\{Z : h(Z) \neq 0\}$, division by $h(Z)$ is valid, yielding

$$a \exp\big(-b\, h(Z)\big) = a' \exp\big(-b'\, h(Z)\big) \qquad \text{a.s. on } \{h(Z) \neq 0\}. \tag{56}$$

Rearranging gives

$$\frac{a}{a'} = \exp\big((b - b')\, h(Z)\big) \qquad \text{a.s. on } \{h(Z) \neq 0\}. \tag{57}$$

The left-hand side is constant in $Z$. If $b \neq b'$, then the right-hand side is a nonconstant function of $h(Z)$. By Assumption 1, the essential range of $h(Z)$ contains at least two distinct nonzero points, so $h(Z)$ takes at least two distinct values on $\{h(Z) \neq 0\}$ and this is impossible. Hence $b = b'$. Substituting back into Eq. (57) gives $a/a' = 1$, so $a = a'$. □

Lemma 1 shows that, in this restricted heteroscedastic family, equality of the induced logit field forces equality of the latent score–scale decomposition itself. Since exact local identifiability is stronger than quotient identifiability, the family is also locally identifiable up to global scaling in the sense of Definition 2.

### D.4 Local statistical non-degeneracy

We next verify that the locally distinct directions remain visible through the Bernoulli observation model. Let

$$p_{a,b}(Z) = \sigma(\eta_{a,b}(Z)), \qquad \eta_{a,b}(Z) = \beta\, a\, h(Z) \exp\big(-b\, h(Z)\big),$$

and let $Y \mid Z \sim \mathrm{Bernoulli}(p_{a,b}(Z))$. The conditional log-likelihood is

$$\ell(a, b; Y, Z) = Y \log p_{a,b}(Z) + (1 - Y) \log(1 - p_{a,b}(Z)).$$

Its score is

$$\nabla_{(a,b)} \ell(a, b; Y, Z) = \big(Y - p_{a,b}(Z)\big)\, \nabla_{(a,b)} \eta_{a,b}(Z), \tag{58}$$

with

$$\frac{\partial \eta_{a,b}(Z)}{\partial a} = \beta\, h(Z) \exp\big(-b\, h(Z)\big), \tag{59}$$

$$\frac{\partial \eta_{a,b}(Z)}{\partial b} = -\beta\, a\, h(Z)^2 \exp\big(-b\, h(Z)\big). \tag{60}$$

Therefore the Fisher information matrix is

$$I(a, b) = \mathbb{E}\big[p_{a,b}(Z)\big(1 - p_{a,b}(Z)\big)\, \nabla \eta_{a,b}(Z) \nabla \eta_{a,b}(Z)^\top\big]. \tag{61}$$

**Lemma 2** (Local Fisher non-degeneracy)**.** *Under Assumption 1, at any point $(a, b)$ with $a \neq 0$, the Fisher information matrix $I(a, b)$ is positive definite.*

*Proof.* For any nonzero vector $v = (v_1, v_2)^\top \in \mathbb{R}^2$,

$$v^\top I(a, b) v = \mathbb{E}\Big[p_{a,b}(Z)\big(1 - p_{a,b}(Z)\big)\, \big(v^\top \nabla \eta_{a,b}(Z)\big)^2\Big]. \tag{62}$$

Because $p_{a,b}(Z) \in (0, 1)$ under the logistic model, the weight $p_{a,b}(Z)(1 - p_{a,b}(Z))$ is strictly positive almost surely. Thus $v^\top I(a, b) v = 0$ can hold only if

$$v^\top \nabla \eta_{a,b}(Z) = 0 \qquad \text{a.s.}$$

Using the derivatives above,

$$v^\top \nabla \eta_{a,b}(Z) = \beta \exp\big(-b\, h(Z)\big)\Big(v_1 h(Z) - a v_2 h(Z)^2\Big). \tag{63}$$

Since $\beta > 0$ and $\exp(-b\, h(Z)) > 0$, this implies

$$h(Z)\big(v_1 - a v_2 h(Z)\big) = 0 \qquad \text{a.s.}$$

On the set $\{Z : h(Z) \neq 0\}$, we therefore have

$$v_1 - a v_2 h(Z) = 0 \qquad \text{a.s.}$$

If $v_2 \neq 0$, then $h(Z) = v_1/(a v_2)$ almost surely on $\{h(Z) \neq 0\}$, contradicting Assumption 1. Hence $v_2 = 0$. Then $v_1 h(Z) = 0$ almost surely, and since $h(Z) \neq 0$ on a set of positive probability by Assumption 1, it follows that $v_1 = 0$. Therefore $v = 0$, proving that $I(a, b)$ is positive definite. □

### D.5 Conclusion for the worked example

We can now summarize the role of this example.

**Corollary 3** (Local identifiability of the worked heteroscedastic family)**.** *Under Assumption 1, the family in Eqs. (51)–(52) is locally identifiable up to global scaling in the sense of Definition 2. Moreover, at any point $(a, b)$ with $a \neq 0$, the induced Bernoulli model is locally non-degenerate for likelihood-based inference.*

*Proof.* Lemma 1 gives local identifiability of the decomposition, which is stronger than local identifiability up to global scaling. Lemma 2 gives local Fisher non-degeneracy. □

Corollary 3 should be read only as an illustrative existence result. It does not claim that the full neural H-DPO implementation is exactly identifiable, nor that arbitrary feature-dependent scale families are locally identifiable from binary preference data without additional structural restrictions. Its role is simply to show that the branch-level identifiability conditions invoked in the main text are realized by a concrete *genuinely heteroscedastic* restricted family, so that the local identifiability claim is not merely heuristic.

*Remark* 1 (An operable design principle for the scale feature). The worked example isolates what makes the score–scale split identifiable: the scale must be driven by a comparison feature $h(Z)$ that genuinely varies across pairs (Assumption 1), so that the score amplitude $a$ and the log-scale slope $b$ leave distinct, non-degenerate imprints on the field $\eta_{a,b}(Z) = \beta\, a\, h(Z) \exp(-b\, h(Z))$. A natural heuristic way to obtain such a feature is to use a response-symmetric quantity that is constant when the two responses receive equal score, for instance an absolute-difference or distance-type feature comparing the two response representations; such features vary with comparison difficulty while remaining invariant to swapping $y_w$ and $y_l$, which is consistent with the permutation-invariant scale-head design used in the experiments (Appendix E). We stress the scope of this principle: it is a design guideline distilled from a restricted single-feature family, not a verification that the full neural head, whose feature span is richer than the $h$ analyzed here, satisfies the corresponding non-degeneracy conditions; establishing those for the deployed head remains open.

## E Implementation Details for H-DPO

This appendix gives the concrete implementation used for H-DPO in the experiments. The theoretical results in the main text concern the heteroscedastic-logit branch as a model class rather than any particular neural parameterization. The implementation below should therefore be read as a practical instantiation of that branch, designed to preserve the key modeling constraint of the method: the comparison-specific scale enters *inside* the preference logit as a divisor rather than appearing only as a loss weight or an additive correction. In particular, neither the specific feature design nor the minibatch-centering trick below is part of the identifiability claim in the main text.

### E.1 Overview

Given a comparison pair $z_i = (x_i, y_w^i, y_l^i)$, H-DPO predicts a positive comparison-specific scale $\tau_i$ and uses it in the conditional preference model

$$P(y_w^i \succ y_l^i \mid x_i) = \sigma\left(\beta\, \frac{\Delta_\pi(x_i, y_w^i, y_l^i)}{\tau_i}\right), \tag{64}$$

where $\Delta_\pi$ is the standard DPO log-ratio score and $\beta$ is fixed to the same global inverse temperature used by the DPO baseline. The scale $\tau_i$ is produced by a lightweight *symmetric ambiguity head* that reads pooled prompt/response representations from the policy model.

The implementation is designed to satisfy three constraints:

1. **positivity:** $\tau_i > 0$ for every pair;
2. **symmetry:** swapping the two responses should not change the predicted scale;
3. **separation from the score channel:** the scale head should model comparison ambiguity rather than preference direction.

### E.2 SCALE-HEAD PARAMETERIZATION

Let $h_w \in \mathbb{R}^H$ and $h_l \in \mathbb{R}^H$ denote pooled completion representations for the two responses in an observed pair, where the subscripts follow the chosen/rejected notation of the training data, and let $h_x \in \mathbb{R}^H$ denote a pooled prompt representation. Although the raw pair is presented in chosen/rejected order, the scale head receives only permutation-invariant transforms of $(h_w, h_l)$ and sign-invariant scalar cues, so its output is unchanged under swapping the two responses.

The scale head takes a pair feature vector $\varphi(z_i) \in \mathbb{R}^{4H+7}$ as input and outputs a scalar local log-scale:

$$s_\phi(z_i) = \mathrm{MLP}_\phi\big(\varphi(z_i)\big), \tag{65}$$

where $\mathrm{MLP}_\phi$ is a two-layer network

$$\mathbb{R}^{4H+7} \;\to\; \mathbb{R}^H \;\to\; \mathbb{R}$$

with a ReLU hidden nonlinearity.

The final scale is parameterized on the log scale as

$$\log \tau_i = b_\phi + \tilde{s}_\phi(z_i), \qquad \tau_i = \exp\big(b_\phi + \tilde{s}_\phi(z_i)\big), \tag{66}$$

where $b_\phi \in \mathbb{R}$ is a learnable global log-scale parameter and $\tilde{s}_\phi(z_i)$ is the centered local log-scale

$$\tilde{s}_\phi(z_i) = s_\phi(z_i) - \mathrm{sg}\left(\frac{1}{B}\sum_{j=1}^{B} s_\phi(z_j)\right). \tag{67}$$

Here $B$ is the minibatch size and $\mathrm{sg}(\cdot)$ denotes stop-gradient.

The purpose of Eq. (67) is to stabilize the decomposition between global and local scale. Since only the ratio $\beta\,\Delta/\tau$ is identified, a global multiplicative change in $\tau$ can be absorbed into the global normalization. In practice, we therefore fix $\beta$ to the standard DPO value and let $b_\phi$ absorb the global scale, while $\tilde{s}_\phi$ captures only relative pairwise variation. This centering is purely an optimization device and is not part of the branch-level model definition.

### E.3 SYMMETRIC PAIR FEATURES

The feature map $\varphi(z_i)$ is constructed to be permutation-invariant with respect to the two responses, so that the scale head models comparison ambiguity rather than directional preference. It consists of a $4H$-dimensional semantic backbone together with seven scalar ambiguity cues.

**Semantic geometry backbone.** The $4H$-dimensional backbone is

$$\big[h_x,\; h_w + h_l,\; |h_w - h_l|,\; h_w \odot h_l\big], \tag{68}$$

where $h_x$ is the pooled prompt representation, $h_w + h_l$ is the response-pair sum, $|h_w - h_l|$ is the elementwise absolute difference, and $h_w \odot h_l$ is the elementwise product. All four blocks are symmetric under swapping $y_w$ and $y_l$.

**Scalar ambiguity cues.** In addition to the semantic backbone, we include seven scalar features:

1. cosine similarity between the two response representations;
2. absolute frozen-reference log-probability margin;
3. frozen-reference binary ambiguity proxy $4p_{\mathrm{ref}}(1 - p_{\mathrm{ref}})$, where $p_{\mathrm{ref}} = \sigma(r_{\mathrm{ref}})$ and $r_{\mathrm{ref}}$ is the frozen-reference log-probability gap between the two responses;
4. absolute gap in prompt–response cosine similarity;
5. absolute gap in prompt–response Euclidean distance;
6. prompt-anchored relative separation of the two responses;
7. normalized response-pair discriminability.

The last two are normalized geometric cues derived from the pooled representations and are included only as auxiliary ambiguity signals.

The frozen-reference cues enter only through sign-invariant transforms, such as absolute margins and ambiguity proxies. They therefore do not provide chosen-versus-rejected directional information to the scale head and do not alter the DPO score channel itself.

**Pooling.** Completion and prompt representations are obtained by masked mean pooling over the final hidden states of the policy model. Completion pooling averages over answer tokens, while prompt pooling averages over the complementary prompt-token positions.

**Optional feature normalization.** The implementation supports optional per-block normalization of the $4H$-dimensional semantic backbone before it is passed to the scale head. In the main experiments this normalization is disabled by default. When enabled, each $H$-dimensional block is normalized separately so that the scale head is less sensitive to base-model-specific representation magnitude.

### E.4 Initialization and trainability

The scale head is attached to the policy model before trainer initialization so that its parameters are included in optimization. The final linear layer of the head is initialized to zero, and the global log-scale parameter $b_\phi$ is initialized at zero. This makes the initial scale exactly constant and equal to one across pairs, so training starts at the homoscedastic DPO baseline.

In implementations that wrap the base model with parameter-efficient fine-tuning modules, the custom H-DPO parameters must remain trainable after wrapping so that the heteroscedastic branch stays active. The real-world experiments in this paper use full-parameter fine-tuning, so no additional PEFT-specific handling is required there.

### E.5 Training objective

For a minibatch $\mathcal{B}$ of size $B$, the implemented training objective is

$$\mathcal{L}^{(\mathcal{B})}_{\text{train}}(\pi, \phi) = \frac{1}{B} \sum_{i \in \mathcal{B}} \left[ -\log \sigma \left( \beta \, \frac{\Delta_\pi(x_i, y_w^i, y_l^i)}{\tau_i} \right) \right] + \lambda_\tau \frac{1}{B} \sum_{i \in \mathcal{B}} \tilde{s}_\phi(z_i)^2, \tag{69}$$

where $\tau_i$ is given by Eq. (66). The regularization term penalizes the centered local log-scale rather than the global log-scale. Its role is not to impose the theory, but to stabilize optimization by discouraging unnecessarily large pair-specific deviations from the homoscedastic baseline.

In the implementation, $\lambda_\tau$ is set to a small positive constant. The regularizer is applied only to the minibatch-centered local component $\tilde{s}_\phi$, so the global scale remains free to absorb the overall normalization.

### E.6 Inside-logit scale correction versus loss reweighting

A central implementation point is that H-DPO modifies the *preference logit* rather than only the empirical loss weight. Concretely, the pair-specific scale is applied as

$$\beta \, \Delta_\pi(z_i) / \tau_i$$

inside the sigmoid. This is the mechanism corresponding to the error-layer interpretation in the main text.

For comparison, the loss-layer control used in the weighted-DPO ablation keeps the same DPO logit $\beta \, \Delta_\pi(z_i)$ and instead multiplies the per-pair loss by a learned weight. The two mechanisms share the same broad pair-adaptive flavor, but they are not equivalent: the former changes the modeled conditional preference probability, whereas the latter changes only the optimization weighting.

### E.7 Ablation-specific implementation variants

Several ablations reuse the same feature pipeline while changing how the head output is used.

**No-$\tau$-branch additive control.** In the additive-head ablation, the head output is injected additively into the preference logit while $\tau$ is pinned to one. This preserves the extra head and its capacity but removes the inside-logit division structure that defines H-DPO.

**Loss-weight control.** In the weighted-DPO control, a separate head with the same input dimensionality and architecture predicts a per-pair loss weight. The weight is squashed to $(0, 1)$ by a sigmoid and then normalized to have batch mean one. This keeps the overall loss scale comparable to DPO while allowing only relative reweighting across pairs.

**Feature masking ablations.** The feature-mask ablations keep the full $4H+7$ input dimensionality but zero out selected feature blocks before the head. This preserves parameter count and isolates the contribution of different feature groups without changing the head architecture.

### E.8 REWARDBENCH SCORING

For RewardBench v1 and v2, responses are scored using the standard DPO implicit reward

$$\hat{r}(x, y) = \beta \sum_t \log \frac{\pi_\theta(y_t \mid x, y_{<t})}{\pi_{\text{ref}}(y_t \mid x, y_{<t})}, \tag{70}$$

summed over answer tokens. A comparison is counted as correct when the chosen response receives a larger implicit reward than the rejected response. This scoring rule is applied identically to all log-ratio-based methods, including H-DPO; SimPO is the sole exception and is scored by its reference-free per-token convention (see Appendix G).

We use the standard implicit reward for comparability across methods and because RewardBench evaluates response ranking under a single-response scoring rule. By contrast, the learned scale in H-DPO is comparison-specific and defined on response pairs rather than on individual responses. Using $\tau_\phi$ directly at evaluation would therefore require a pairwise scoring rule rather than the standard single-response implicit reward protocol.

## F SYNTHETIC EXPERIMENT DETAILS

This appendix gives the full protocol for the synthetic study summarized in Section 5.1. The purpose of the synthetic experiment is to test the branch-level predictions of Section 4 in a setting where the latent utility gap $\Delta^\star$, the comparison-specific scale $\tau^\star$, and the true conditional preference probability $p^\star$ are all known. Real preference benchmarks cannot provide this information, so they cannot directly distinguish recovery of the true utility-gap object from recovery of the distorted surrogate $\Delta^\star/\tau^\star$.

### F.1 DATA-GENERATING PROCESS

For each comparison $i$, we independently draw two feature blocks

$$z_i^\Delta,\ z_i^\tau \sim \mathcal{N}(0, I_8),$$

and define

$$\Delta_i^\star = {w_\Delta^\star}^\top z_i^\Delta, \tag{71}$$

$$\tau_i^\star = \exp\big(\alpha\, {w_\tau^\star}^\top z_i^\tau\big), \tag{72}$$

where $w_\Delta^\star$ and $w_\tau^\star$ are independent unit-norm direction vectors sampled once per seed. The observed binary label is then generated as

$$r_i \sim \text{Bernoulli}(p_i^\star), \qquad p_i^\star = \sigma(\beta\, \Delta_i^\star/\tau_i^\star)\,, \qquad \beta = 1. \tag{73}$$

The model input is the concatenated feature vector

$$z_i = [z_i^\Delta; z_i^\tau] \in \mathbb{R}^{16}.$$

Thus all methods receive the same information. Any failure of homoscedastic DPO to recover $\Delta^\star$ is therefore due to model misspecification rather than missing covariates.

We sweep the heterogeneity strength over

$$\alpha \in \{0, 0.25, 0.5, 0.75, 1.0\},$$

where $\alpha = 0$ is the homoscedastic degenerate case and larger $\alpha$ induces larger variance in $\log \tau_i^\star$. For each $\alpha$, we generate $50{,}000$ training pairs, $10{,}000$ validation pairs, and $10{,}000$ test pairs, and repeat the experiment over five random seeds.

### F.2 Models and training configuration

We compare three methods:

1. **DPO**: a homoscedastic score model implemented by the shared score network;
2. **H-DPO**: the same score network augmented with a heteroscedastic scale head;
3. **W-DPO**: an oracle weighted-DPO control that uses the true $1/\tau_i^\star$ only as a per-pair loss weight while keeping the comparison channel homoscedastic.

All three methods use the same score network: a three-layer MLP with hidden size $128$ and ReLU activations. H-DPO uses an additional scale head of the same architecture. The final linear layer of the H-DPO scale head is zero-initialized, and the global log-scale is initialized at zero. Consequently, the initial local log-scale is zero and the initial scale is $\tau_i = 1$ for all pairs, so H-DPO starts from the homoscedastic DPO baseline.

All methods are trained with Adam using learning rate $10^{-3}$, batch size $256$, cosine decay, gradient clipping at global norm $1.0$, and early stopping on validation NLL with patience $10$. The global inverse temperature is fixed to $\beta = 1$ for all methods.

The only additional H-DPO hyperparameter is the scale regularization coefficient $\lambda_\tau$. To avoid per-condition tuning, we select a single global value once using validation NLL averaged over the moderate heterogeneity settings $\alpha \in \{0.25, 0.5, 0.75\}$, and then keep it fixed for the full sweep. The selected value is

$$\lambda_\tau = 10^{-3}.$$

For W-DPO, the oracle weight is

$$w_i \propto 1/\tau_i^\star,$$

normalized to have mean one within the training set so that weighting changes only the relative contribution of pairs rather than the overall loss scale. This choice mirrors the inside-logit divisor as closely as a loss-layer control can, while remaining a purely loss-layer intervention.

### F.3 Metrics

Each reported metric corresponds directly to a theoretical quantity from Section 4.

**Utility-gap recovery.** We measure recovery of the true latent utility gap by

$$\rho_\Delta = \mathrm{Corr}\big(\hat{\Delta}_i, \Delta_i^\star\big),$$

where $\hat{\Delta}_i$ is the learned score output of the model and $\mathrm{Corr}$ denotes Pearson correlation. Pearson correlation is used here as a scale-free recovery metric: it is invariant to global rescaling, matching the branch-level scaling ambiguity under which the score is identified.

**Distorted-target alignment.** To test Corollary 1, we also compute

$$\rho_{\Delta/\tau} = \mathrm{Corr}\big(\hat{\Delta}_i, \Delta_i^\star/\tau_i^\star\big).$$

The signed distortion diagnostic is then

$$G_{\mathrm{dist}} = \rho_{\Delta/\tau} - \rho_\Delta.$$

Positive values indicate stronger alignment with the distorted surrogate $\Delta^\star/\tau^\star$ than with the true utility gap $\Delta^\star$.

**Test negative log-likelihood.** We report test NLL to verify that differences in $\rho_\Delta$ are not simply due to one method fitting the observed labels better than another.

**Scale recovery.** For H-DPO, we additionally report

$$\rho_{\log \tau} = \mathrm{Corr}\big(\log \hat{\tau}_i - \overline{\log \hat{\tau}},\ \log \tau_i^\star - \overline{\log \tau^\star}\big),$$

where centering removes the global multiplicative ambiguity and isolates recovery of the comparison-specific scale structure. When $\tau_i^\star$ is constant (here $\alpha = 0$), this correlation is undefined; in Table 2 it is reported as $0.000$ by convention.

### F.4 Interpretation of the synthetic controls

The synthetic experiment is designed so that each method differs only in specification, not in observed information. DPO and H-DPO receive the same input features, and W-DPO is given the strongest possible loss-layer control by using the true $1/\tau_i^\star$ as its weight. The role of W-DPO is therefore diagnostic: if oracle reweighting still tracks DPO rather than H-DPO, then the difference between H-DPO and DPO cannot be attributed merely to pair-adaptive weighting.

The synthetic results in Table 2 support exactly this interpretation. As $\alpha$ increases, DPO increasingly aligns with the distorted target $\Delta^\star/\tau^\star$, while H-DPO maintains substantially higher alignment with the true utility gap $\Delta^\star$. At the same time, test NLL remains nearly identical across methods, showing that the main difference is not label fit but which latent quantity is recovered.

The synthetic setup also helps rule out a pure capacity explanation. Although H-DPO has an additional scale head, DPO already matches H-DPO closely in test NLL. The advantage of H-DPO is therefore not that it predicts the binary labels much better, but that it allocates variation between score and scale in a way that matches the data-generating branch.

### F.5 Reproducibility

The full synthetic sweep is reproduced by running the released synthetic experiment code with the fixed seed set $\{0, 1, 2, 3, 4\}$ and the five heterogeneity settings $\alpha \in \{0, 0.25, 0.5, 0.75, 1.0\}$. The released code also contains the global selection routine for $\lambda_\tau$ described above, and all reported numbers are aggregated over the fixed five-seed protocol.

### F.6 Paired seed-level robustness of the distortion signature

Table 2 aggregates the synthetic sweep as a mean over the five seeds. Because the same five seeds are run for every method at every heterogeneity level, the H-DPO-versus-DPO comparison is *paired within seed*, so we can also ask how stable the recovered-quantity gap is across independent training runs rather than only across evaluation resamples. Table 5 reports this paired view. For each $\alpha$ and seed we compute the alignment advantage $\Delta\rho_\Delta = \rho_\Delta^{\text{H-DPO}} - \rho_\Delta^{\text{DPO}}$, the per-seed margin by which the H-DPO score stays closer to the true utility gap $\Delta^\star$ than the DPO score does, together with the distortion gap $G_{\text{dist}}$ of each method.

At $\alpha = 0$ the generative model is homoscedastic and the two methods coincide, as the theory requires: $\Delta\rho_\Delta = -0.0008 \pm 0.0014$, positive in only $2/5$ seeds, paired $t = -1.24$ on four degrees of freedom (not significant). For every $\alpha > 0$, the advantage is positive in all five seeds ($5/5$). The paired $t$ statistics are uniformly large ($t \geq 14$), and the exact one-sided sign test at $5/5$ gives $p = 0.03125$. Given the small number of seeds, we report these tests only as descriptive summaries of run-level consistency. The advantage grows monotonically with heterogeneity, from $+0.0145 \pm 0.0023$ at $\alpha = 0.25$ to $+0.1704 \pm 0.0136$ at $\alpha = 1.0$. The distortion gap behaves dually. Once $\alpha \geq 0.5$, DPO's $G_{\text{dist}}$ is positive in every seed (drift toward $\Delta^\star/\tau^\star$) while H-DPO's is negative in every seed (retention of the $\Delta^\star$ side), so the qualitative signature is not carried by a single favorable run.

**Scope.** This paired analysis establishes *training-run* robustness for the synthetic branch, where five independent runs per setting are available. It does not by itself establish run-to-run robustness for the real-world models, which are trained with a single seed per base model (see Appendix G).

Table 5: Paired seed-level robustness of the synthetic distortion signature ($n = 5$ seeds, $\{0, 1, 2, 3, 4\}$). $\Delta\rho_\Delta = \rho_\Delta^{\text{H-DPO}} - \rho_\Delta^{\text{DPO}}$ is the per-seed alignment advantage toward the true utility gap $\Delta^\star$; "pos." counts seeds with $\Delta\rho_\Delta > 0$; $t$ is the paired $t$ statistic on four degrees of freedom. $G_{\text{dist}}$ is each method's signed distortion gap: positive drifts toward $\Delta^\star/\tau^\star$, negative stays on the $\Delta^\star$ side. Entries are mean±sd over seeds.

| $\alpha$ | $\Delta\rho_\Delta$ | pos. | paired $t$ | DPO $G_{\text{dist}}$ | H-DPO $G_{\text{dist}}$ |
|---|---|---|---|---|---|
| 0.00 | $-0.0008 \pm 0.0014$ | 2/5 | $-1.24$ | $+0.0000 \pm 0.0000$ | $+0.0000 \pm 0.0000$ |
| 0.25 | $+0.0145 \pm 0.0023$ | 5/5 | $+14.00$ | $+0.0214 \pm 0.0022$ | $-0.0031 \pm 0.0040$ |
| 0.50 | $+0.0643 \pm 0.0075$ | 5/5 | $+19.10$ | $+0.0706 \pm 0.0101$ | $-0.0378 \pm 0.0083$ |
| 0.75 | $+0.1149 \pm 0.0121$ | 5/5 | $+21.31$ | $+0.1078 \pm 0.0178$ | $-0.0935 \pm 0.0281$ |
| 1.00 | $+0.1704 \pm 0.0136$ | 5/5 | $+28.10$ | $+0.1073 \pm 0.0343$ | $-0.2045 \pm 0.0446$ |

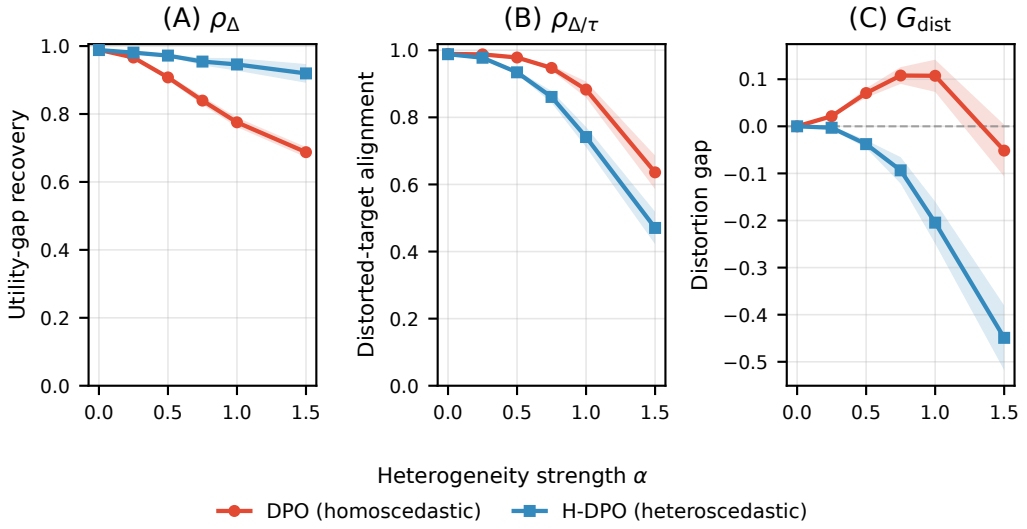



Figure 2: Synthetic verification of the semantic misspecification prediction. As heterogeneity increases, DPO aligns increasingly with the distorted target $\Delta^\star/\tau^\star$, whereas H-DPO maintains substantially stronger alignment with the true utility gap $\Delta^\star$. Shaded regions denote $\pm 1$ standard deviation over five seeds.

What the real-world experiments provide instead is *cross-base* replication of the same preference-discrimination direction. Across the three base-model families, all six RewardBench contrasts in Table 13 favor H-DPO (from $+0.05$ to $+1.78$ percentage points), four of the six are significant under the paired bootstrap, and the two Mistral-7B contrasts fall inside a $\pm 1$-point band around zero, which we use only as a descriptive margin rather than as a formal equivalence test. Cross-base consistency and cross-seed consistency are distinct forms of evidence, and for the real-world rows we claim only the former.

### F.7 Additional synthetic visualization

Figure 2 provides a compact visualization of this pattern.

### F.8 Ranking invariance versus magnitude-sensitive selection

Pairwise preference discrimination, including RewardBench, consumes only the *sign* of the score difference within each comparison. Best-of-N selection and cross-prompt aggregation instead consume score *magnitudes*. The population theory separates these two channels: because $\tau^\star > 0$, the distorted target $\Delta^\star/\tau^\star$ is sign-identical to $\Delta^\star$ pair-wise, so population pairwise ranking accuracy is

Table 6: Ranking invariance versus magnitude-sensitive best-of-N selection on the synthetic branch (mean±sd over five seeds). Pairwise ranking accuracy and rank AUC are statistically indistinguishable across DPO, H-DPO, and oracle W-DPO at every $\alpha$, confirming that the distortion is ranking-invariant. Magnitude-sensitive best-of-N accuracy and regret separate sharply as $\alpha$ grows, with H-DPO retaining high selection accuracy and low regret while DPO and W-DPO degrade; this is the channel where utility-faithful semantics pays off and which pairwise discrimination does not consume.

| $\alpha$ | Method | rank acc | rank AUC | BoN-8 acc | BoN-8 regret |
|---|---|---|---|---|---|
| 0.00 | DPO | $0.955 \pm 0.002$ | $0.994 \pm 0.001$ | $0.887 \pm 0.006$ | $0.015 \pm 0.002$ |
| 0.00 | H-DPO | $0.954 \pm 0.006$ | $0.994 \pm 0.001$ | $0.886 \pm 0.018$ | $0.016 \pm 0.005$ |
| 0.00 | W-DPO | $0.955 \pm 0.004$ | $0.994 \pm 0.001$ | $0.890 \pm 0.013$ | $0.015 \pm 0.004$ |
| 0.25 | DPO | $0.955 \pm 0.004$ | $0.993 \pm 0.001$ | $0.787 \pm 0.008$ | $0.056 \pm 0.005$ |
| 0.25 | H-DPO | $0.955 \pm 0.003$ | $0.994 \pm 0.001$ | $0.841 \pm 0.017$ | $0.031 \pm 0.006$ |
| 0.25 | W-DPO | $0.954 \pm 0.004$ | $0.993 \pm 0.001$ | $0.783 \pm 0.010$ | $0.058 \pm 0.005$ |
| 0.50 | DPO | $0.952 \pm 0.004$ | $0.993 \pm 0.001$ | $0.679 \pm 0.008$ | $0.129 \pm 0.005$ |
| 0.50 | H-DPO | $0.957 \pm 0.002$ | $0.994 \pm 0.001$ | $0.804 \pm 0.008$ | $0.047 \pm 0.004$ |
| 0.50 | W-DPO | $0.952 \pm 0.003$ | $0.992 \pm 0.001$ | $0.677 \pm 0.020$ | $0.132 \pm 0.013$ |
| 0.75 | DPO | $0.941 \pm 0.009$ | $0.990 \pm 0.002$ | $0.611 \pm 0.013$ | $0.189 \pm 0.012$ |
| 0.75 | H-DPO | $0.957 \pm 0.005$ | $0.994 \pm 0.001$ | $0.756 \pm 0.036$ | $0.075 \pm 0.022$ |
| 0.75 | W-DPO | $0.952 \pm 0.008$ | $0.992 \pm 0.001$ | $0.615 \pm 0.015$ | $0.188 \pm 0.013$ |
| 1.00 | DPO | $0.937 \pm 0.008$ | $0.987 \pm 0.003$ | $0.556 \pm 0.011$ | $0.246 \pm 0.012$ |
| 1.00 | H-DPO | $0.954 \pm 0.008$ | $0.993 \pm 0.002$ | $0.742 \pm 0.035$ | $0.084 \pm 0.025$ |
| 1.00 | W-DPO | $0.941 \pm 0.013$ | $0.988 \pm 0.004$ | $0.567 \pm 0.009$ | $0.239 \pm 0.011$ |

unchanged by the distortion, whereas magnitudes are not. Pairwise discrimination therefore cannot by itself attest to semantic recovery; magnitude-sensitive tasks can.

We test this separation on the synthetic branch. For each $(\alpha, \text{seed})$ and method we compute, on the test set, the pairwise ranking accuracy (fraction of pairs with $|\Delta^\star| \geq \varepsilon$ on which $\text{sign}(\hat{\Delta}) = \text{sign}(\Delta^\star)$) and the rank AUC (Mann-Whitney) for the class $\Delta^\star > 0$. For magnitude sensitivity we draw $4{,}000$ groups of $N \in \{4, 8, 16\}$ test comparisons per $(\alpha, \text{seed})$ under a fixed seed, select the member with the largest predicted score, and record the selection accuracy (agreement with the largest true gap) and the regret $\mathbb{E}[\max \Delta^\star - \Delta^\star_{\text{selected}}]$.

Table 6 confirms the predicted separation. Ranking accuracy and rank AUC overlap within one standard deviation across all three methods at every $\alpha$: pairwise discrimination is blind to the distortion, exactly as the sign-identity of $\Delta^\star/\tau^\star$ implies. Best-of-N selection, by contrast, separates sharply once $\alpha > 0$: at $\alpha = 1.0$ the BoN-8 accuracy is $0.556$ for DPO and $0.567$ for oracle W-DPO but $0.742$ for H-DPO, with regret $0.246$/$0.239$ versus $0.084$. Oracle W-DPO tracks DPO on both magnitude metrics, consistent with Proposition 4: loss-layer reweighting changes neither the ranking-invariance nor the magnitude distortion. We therefore interpret the real-world RewardBench gains as consistent with a finite-sample optimization channel in the deployed model class, which we isolate with a dedicated experiment next, rather than as direct evidence of population semantic recovery; and we locate the operational value of utility-faithful semantics in magnitude-consuming downstream use such as best-of-N selection.

Two candidate mechanisms could produce such a finite-sample gain. Rather than assert them, we isolate each with a dedicated synthetic experiment built on the existing branch protocol (same DGP, same MLP, five paired seeds per $\alpha$).

**The finite-sample optimization channel is verified.** The heteroscedastic per-pair loss $-\log \sigma(\beta\Delta_i/\tau_i)$ has score gradient $-(\beta/\tau_i)\,\sigma(-\beta\Delta_i/\tau_i)$, so a comparison the model reads as more ambiguous (larger $\tau_i$) receives a proportionally smaller update. This factor is closed-form, so we measure it directly at the trained model's own test-set predictions without any backpropagation. Table 7 reports the association between the per-pair score-gradient magnitude $g_i$ and ambiguity. The homoscedastic DPO gradient $\beta\,\sigma(-\beta\Delta_i)$ carries no per-pair ambiguity modulation: its Spearman correlation with the true scale $\tau^\star$ is indistinguishable from zero at every heterogeneity level ($|\rho| \leq 0.02$), and the ratio of its mean gradient in the highest-$\tau^\star$ tercile to the lowest is $1.01 \pm 0.03$ at $\alpha = 1.0$. The H-DPO gradient behaves exactly as the mechanism predicts. Its correlation with

Table 7: Isolating the finite-sample optimization channel on the synthetic branch (mean±sd over five seeds). For each method we evaluate the closed-form per-pair score-gradient magnitude $g_i$ at the trained model's test-set predictions and report its Spearman correlation with the true scale $\tau^\star$ (and, for H-DPO, with the model's learned scale $\tau_\phi$), together with the ratio of the mean gradient in the highest-$\tau^\star$ tercile to that in the lowest. DPO's gradient is flat in ambiguity (correlation near zero, tercile ratio near one): a single global $\beta$ cannot modulate the update per pair. H-DPO's gradient is increasingly negatively tied to ambiguity as $\alpha$ grows, so the most ambiguous, most label-noisy comparisons receive the smallest updates, which is the mechanism the theory predicts. $\alpha = 0$ is omitted because $\tau^\star \equiv 1$ leaves no ambiguity to modulate.

| $\alpha$ | DPO $\rho_S(g, \tau^\star)$ | H-DPO $\rho_S(g, \tau^\star)$ | H-DPO $\rho_S(g, \tau_\phi)$ | H-DPO hi/lo |
|---|---|---|---|---|
| 0.25 | $+0.010 \pm 0.020$ | $-0.156 \pm 0.039$ | $-0.239 \pm 0.082$ | $0.809 \pm 0.035$ |
| 0.50 | $+0.004 \pm 0.022$ | $-0.308 \pm 0.036$ | $-0.349 \pm 0.049$ | $0.572 \pm 0.042$ |
| 0.75 | $-0.005 \pm 0.013$ | $-0.362 \pm 0.063$ | $-0.388 \pm 0.065$ | $0.423 \pm 0.073$ |
| 1.00 | $+0.002 \pm 0.015$ | $-0.398 \pm 0.057$ | $-0.414 \pm 0.062$ | $0.280 \pm 0.049$ |

$\tau^\star$ grows monotonically more negative with heterogeneity (Spearman $-0.156 \to -0.398$ from $\alpha = 0.25$ to 1.0), as does its correlation with the model's own learned scale $\tau_\phi$ ($-0.239 \to -0.414$), and the highest-ambiguity tercile receives a progressively smaller update, the high/low gradient ratio falling from $0.809$ to $0.280$. At $\alpha = 1.0$ the most-ambiguous tercile receives a mean score update of $0.292 \pm 0.023$ against $1.053 \pm 0.084$ for the least-ambiguous tercile, a factor of about 3.6, while DPO's tercile means are flat at $0.50$. Because the most ambiguous comparisons are also the ones closest to a coin flip, and therefore the most label-noisy, this factor acts as an implicit per-pair down-weighting of noisy comparisons that a single global $\beta$ applies uniformly and cannot modulate per pair. Consistent with this being the active finite-sample mechanism, H-DPO also incurs fewer sign errors than DPO precisely on the hardest, smallest-margin comparisons once heterogeneity is nonzero (small-margin sign error $0.134 \pm 0.019$ versus $0.169 \pm 0.017$ at $\alpha = 1.0$), even though Table 6 shows the two are tied on aggregate ranking accuracy; the benefit is concentrated near the decision boundary, which is where the down-weighting of coin-flip comparisons acts.

**The restricted-class projection channel is not reproduced, so we do not claim it.** The second candidate is a projection effect: the deployed score family is separable, $\Delta_\theta(Z) = f_\theta(x, y_w) - f_\theta(x, y_l)$, and need not contain the distorted target $\Delta^\star/\tau^\star$; the homoscedastic fit then projects that target onto the family, which could in principle reorder pairs whose margin is small relative to the approximation error. We built a purpose-designed control for exactly this restriction. The utility gap is additively separable, $\Delta^\star = a(u_w) - a(u_l)$, so it lies in the family, while the ambiguity carries a non-separable interaction, $\tau^\star = \exp\big(\alpha\, c(u_w)\, c(u_l)\big)$, so $\Delta^\star/\tau^\star$ does not. The restriction bites as intended: the homoscedastic difference-form fit cannot track the distorted target, and its correlation with $\Delta^\star/\tau^\star$ collapses from $0.948 \pm 0.014$ to $0.110 \pm 0.081$ as $\alpha$ goes from $0.25$ to 1.0. Yet the projection does not reorder. The same fit still recovers the utility gap ($\rho_\Delta \approx 0.994$), and its pairwise ranking accuracy against $\mathrm{sign}(\Delta^\star)$ is statistically indistinguishable from H-DPO's at every heterogeneity level ($0.971 \pm 0.003$ versus $0.975 \pm 0.004$ at $\alpha = 1.0$, over five seeds). The projected target remains monotone in $\Delta^\star$, so this control reproduces the ranking-invariance of Corollary 2 rather than a reordering. We therefore do not attribute the real-world gains to restricted-class projection: within our controlled synthetic DGP the projection preserves pairwise ranking, and any reordering in the deployed model class would be an approximation-error effect that our experiments do not exhibit and that we do not claim as verified.

The loss-weight control of Section 5.3 shares only the down-weighting aspect of the verified finite-sample optimization channel: its conditional comparison channel remains homoscedastic, so its population target is still the distorted $\Delta^\star/\tau^\star$ (Proposition 4), and its empirical degradation relative to full H-DPO indicates that generic loss down-weighting alone is not the active ingredient; the active ingredient is the scale entering the conditional preference model itself.

**Why we do not run a real-model best-of-N sweep.** A real-model best-of-N sweep would be the most direct real-world probe of the magnitude channel, but two considerations make it unnecessary for the claim we make. First, the within-model magnitude channel is already demonstrated on the synthetic branch, where the latent scale is known and best-of-N selection separates sharply (Table 6);

a real-model sweep cannot observe the latent scale and would therefore add a behavioral correlate rather than new identification. Second, RewardBench v2 was itself constructed and validated by Malik et al. (2026) precisely by quantifying how benchmark performance correlates with downstream use of the models, including inference-time scaling algorithms such as best-of-N sampling and RLHF training algorithms such as proximal policy optimization. Our RewardBench v2 discrimination results therefore inherit that externally validated predictive validity: strong RewardBench v2 discrimination is evidence of downstream best-of-N and RLHF utility at the cross-model level. We stress the scope of this inheritance. It is a cross-model predictive-validity argument, not a within-model magnitude measurement, and it complements rather than replaces the synthetic within-model best-of-N evidence above.

**Centering variant.** The synthetic proxy centers the local log-scale with a stop-gradient on the batch mean, matching the centering used by the deployed trainer (Section 3.3 and Eq. 67). As a robustness check, rerunning the analysis with a fully differentiable zero-mean projection, which shares the same forward value of $\tau$ at every step and differs only in the gradient routed to the scale head, reproduces the same qualitative pattern (ranking metrics tied, best-of-N diverging with $\alpha$).

# G Full Real-World Experiment Details

This appendix gives the full protocol for the real-world experiments summarized in Section 5.2. We first describe the training data, shared training configuration, and method-specific hyperparameter protocol, and then report the complete benchmark results together with the paired H-DPO-versus-DPO comparisons used in the main text. The focus is on the choices most relevant to fairness, comparability, reproducibility, and interpretation.

## G.1 Training data

All real-world preference-optimization experiments are trained on UltraFeedback-style pairwise preference data aligned to the corresponding base-model family. We use:

- **Llama-3.1-8B-Instruct:** `princeton-nlp/llama3-ultrafeedback`,
- **Gemma-2-9B-it:** `princeton-nlp/gemma2-ultrafeedback-armorm`,
- **Mistral-7B-Instruct-v0.3:** `mistral-instruct-ultrafeedback` from the same dataset collection as the entries above.

Each dataset is used in standard pairwise chosen/rejected format. In our local preprocessing, examples are stored as JSONL records with fields {`messages`, `rejected_messages`}, each representing a two-turn user/assistant dialogue. The resulting train/test sizes are:

- **Llama-3.1-8B-Instruct:** 59,868 train / 1,961 test,
- **Gemma-2-9B-it:** 59,008 train / 1,922 test,
- **Mistral-7B-Instruct-v0.3:** 59,720 train / 2,994 test.

This base-specific data choice keeps the training distribution aligned with the tokenizer, prompt template, and instruction-tuning conventions of each model family while preserving a common binary pairwise-preference setting across methods. No method receives additional supervision such as user identities, ties, or graded labels.

NS-DPO is the only method that uses a dataset variant with an additional field: when `loss_type=ns_dpo`, the training script switches to a corresponding `*_ns.jsonl` file containing timestamps required by the temporal weighting scheme.

## G.2 Shared training configuration

All methods are trained in the same framework, `ms-swift` (`swift rlhf`), using *full-parameter fine-tuning*. Training uses bf16 throughout, with no quantization. For DPO-style methods, the reference model is the frozen initial checkpoint of the same base model, i.e., `ref_model = model`, which is the standard full-parameter DPO setup.

The following settings are held fixed across all three base-model families unless a method definition requires otherwise:

- **fine-tuning mode:** full-parameter fine-tuning;
- **precision:** bf16;
- **epochs:** 1;
- **maximum sequence length:** 2048;
- **global batch size:** 128;
- **learning-rate schedule:** cosine decay with warmup ratio 0.1;
- **optimizer stack:** DeepSpeed ZeRO-2 distributed training.

All main experiments are run on 8 AMD MI308X GPUs under ROCm.

### G.3 Base-specific training differences

Three training settings differ across base-model families and should be read as part of the protocol rather than as method-specific tuning (Table 8):

Table 8: Base-specific training settings that differ across the three model families. These are part of the shared protocol rather than method-specific tuning.

| | **Llama-3.1-8B** | **Mistral-7B-v0.3** | **Gemma-2-9B-it** |
|---|---|---|---|
| Learning rate | $1 \times 10^{-6}$ | $5 \times 10^{-7}$ | $8 \times 10^{-7}$ |
| Per-device batch $\times$ grad. accumulation | $16 \times 1$ | $16 \times 1$ | $8 \times 2$ |
| Attention implementation | sdpa (default) | sdpa (default) | flash attention |

The Gemma-2 batch split differs for memory reasons rather than for optimization-budget reasons. Gemma-2 has a substantially larger vocabulary, so using per-device batch size 16 caused out-of-memory failures in our setup. We therefore use $8 \times 2$ for Gemma-2, which preserves the same global batch size of 128 and hence keeps the overall training budget comparable across bases.

The attention implementation also differs for a model-specific reason. Gemma-2 uses softcapping in its attention/logit stack, and the default sdpa path in our setup does not preserve this behavior reliably for reward-based log-probability evaluation. We therefore use flash attention for Gemma-2 in training and the corresponding `flash_attention_2` path in evaluation. Llama and Mistral do not use this mechanism, so the default sdpa path is used for both training and evaluation.

### G.4 Method families and hyperparameter protocol

All DPO-family methods are trained through the same `swift rlhf` interface with `rlhf_type=dpo` and a method-specific `loss_type`. SimPO is trained separately through `rlhf_type=simpo`. The main compared methods are DPO, IPO, SimPO, $\beta$-DPO, $\varepsilon$-DPO, MixDPO, EM-DPO, NS-DPO, and H-DPO, matching the set reported in the main paper.

**DPO-family methods.** The effective tuning parameters are:

- **DPO** (`loss_type=sigmoid`): $\beta \in \{0.01, 0.05, 0.1\}$.
- **IPO** (`loss_type=ipo`): $\tau \in \{0.01, 0.1, 0.5, 1.0\}$, passed through the `beta` argument in the implementation.
- **$\beta$-DPO** (`loss_type=beta_dpo`): `base_beta` $\in \{0.05, 0.1, 0.2\}$. In this implementation, the externally passed $\beta$ does not affect the loss itself; the effective control parameter is `base_beta`.
- **$\varepsilon$-DPO** (`loss_type=epsilon_dpo`): $\beta$ together with $\varepsilon$, where $\varepsilon$ is kept at its default value 0.01 and the effective $\beta$ is adaptively annealed during training according to the implementation.

- **EM-DPO** (`loss_type=em_dpo`): $\beta \in \{0.01, 0.05, 0.1\}$.
- **MixDPO** (`loss_type=mix_dpo`): no external $\beta$ tuning; the effective $\beta$ is learned by the model.
- **NS-DPO** (`loss_type=ns_dpo`): $\beta \in \{0.01, 0.05, 0.1\}$, using the timestamped dataset variant described above.
- **H-DPO** (`loss_type=hdpo`): $\beta$ together with the scale regularization coefficient $\lambda_\tau$, denoted `HDPO_TAU_REG` in the implementation.

**H-DPO scale-regularization tuning.** The H-DPO scale regularization coefficient is tuned separately for each base model. The searched grids and selected values are:

- **Llama-3.1-8B-Instruct:** $\{0, 10^{-4}, 5 \times 10^{-4}, 10^{-3}, 5 \times 10^{-3}, 0.01, 0.05, 0.1, 0.5, 1\}$, selected $\lambda_\tau = 0.1$;
- **Gemma-2-9B-it:** $\{0, 0.5, 1, 1.5, 2\}$, selected $\lambda_\tau = 0.5$;
- **Mistral-7B-Instruct-v0.3:** $\{0, 0.1, 0.2, 0.3, 0.5\}$, selected $\lambda_\tau = 0.5$.

These values are therefore *per-base tuned*; there is no single globally recommended $\lambda_\tau$ across model families. These ranges were chosen after short pilot runs to cover the stable regime for each base model and should be read as practical per-base tuning ranges rather than as a matched hyperparameter-budget benchmark. More broadly, all methods were tuned within the ranges supported by their standard implementations or official configurations; we did not attempt exhaustive equal-budget hyperparameter optimization across all methods.

**Baselines not included.** uDPO and CAPO are not included as numerical baselines because they target different forms of heterogeneity and supervision regimes from the comparison-specific observation-scale branch isolated here: uDPO penalizes reward uncertainty, and CAPO targets confidence-aware multilingual preference optimization. Reproducing them faithfully would additionally require uncertainty estimates or multilingual supervision outside our single-judge English pairwise protocol. We therefore discuss both qualitatively in the related-work narrative rather than as head-to-head numbers.

**SimPO.** SimPO is trained separately with `rlhf_type=simpo`, no reference model, and `cpo_alpha` $= 0$, following the standard SimPO convention of omitting the auxiliary NLL term. In the `ms-swift` interface, the argument `simpo_gamma` is the *absolute* $\gamma$, whereas official SimPO configurations are often reported in terms of the ratio $\gamma/\beta$. We therefore convert official settings by $\gamma = \beta \times (\gamma/\beta)$.

For Mistral-7B-Instruct-v0.3 and Gemma-2-9B-it, we use the official per-base SimPO settings:

- **Mistral-7B-Instruct-v0.3:** $\beta = 2.5$, $\gamma/\beta = 0.1$, hence $\gamma = 0.25$;
- **Gemma-2-9B-it:** $\beta = 10$, $\gamma/\beta = 0.5$, hence $\gamma = 5.0$.

For Llama-3.1-8B-Instruct, we performed a small search over $\beta \in \{2, 4\}$ and $\gamma \in \{1.0, 1.2, 1.4\}$, and additionally evaluated $(\beta, \gamma) = (10, 3.0)$. The main table uses $(\beta, \gamma) = (10, 3.0)$, which performed best in our runs.

### G.5 Evaluation protocol

All runs are evaluated with a single judge, `gpt-4.1`, at temperature 0. Generation is greedy (`do_sample=false`) with `max_new_tokens` $= 4096$ in bf16. This protocol is held fixed across methods and base models.

**Benchmark versions and judge deviation.** AlpacaEval is evaluated with the length-controlled AlpacaEval 2.0 pipeline, and Arena-Hard with `arena-hard-auto-v0.1` (500 prompts, `gpt-4-0314` as the baseline model, two-game position-swapped comparisons feeding the style-controlled win rate and the headline Elo). Our single judge is `gpt-4.1`, whereas the official AlpacaEval 2.0 and Arena-Hard leaderboards use `gpt-4-1106-preview` (GPT-4-Turbo) as judge.

Table 9: Preference-discrimination axis: RewardBench Overall with 95% bootstrap intervals over benchmark prompts ($B = 10^4$). The TIES subset of RewardBench v2 is held fixed in the resampling because it is aggregated by the official `process_single_model` formula and carries no per-prompt score. †SimPO is reference-free and is therefore not directly commensurable with the log-ratio rows.

| | Llama-3.1-8B-Instruct | | Gemma-2-9B-it | | Mistral-7B-Instruct-v0.3 | |
|---|---|---|---|---|---|---|
| Method | RB v1 | RB v2 | RB v1 | RB v2 | RB v1 | RB v2 |
| *Homoscedastic references* | | | | | | |
| DPO | 75.76 [74.07, 77.48] | 46.32 [44.26, 48.40] | 63.60 [61.86, 65.35] | 36.19 [34.15, 38.33] | 68.00 [66.19, 69.81] | 40.87 [38.77, 42.94] |
| IPO | 69.47 [67.68, 71.24] | 44.67 [42.62, 46.69] | 54.80 [53.10, 56.51] | 28.76 [26.71, 30.82] | 62.98 [61.26, 64.64] | 37.65 [35.50, 39.82] |
| SimPO† | 72.25 [70.74, 73.72] | 47.90 [45.75, 50.01] | 61.37 [59.72, 63.04] | 29.84 [27.79, 31.88] | 66.04 [64.54, 67.53] | 37.98 [35.93, 40.04] |
| *Loss-layer (*L-X*)* | | | | | | |
| $\beta$-DPO | 72.83 [71.05, 74.55] | 41.09 [39.06, 43.15] | 55.37 [53.48, 57.29] | 25.13 [23.21, 27.06] | 60.21 [58.21, 62.21] | 34.61 [32.58, 36.68] |
| $\varepsilon$-DPO | 75.25 [73.53, 76.97] | 47.23 [45.20, 49.29] | 62.30 [60.63, 64.03] | 34.96 [32.89, 37.07] | 68.13 [66.32, 69.94] | 42.12 [40.07, 44.21] |
| *Utility-side (*L-V*)* | | | | | | |
| MixDPO | 76.43 [74.81, 78.09] | 40.16 [38.16, 42.21] | 60.86 [59.02, 62.71] | 30.46 [28.56, 32.41] | 65.83 [63.88, 67.73] | 38.75 [36.72, 40.81] |
| EM-DPO | 74.91 [73.18, 76.58] | 46.41 [44.35, 48.52] | 63.93 [62.16, 65.65] | 36.50 [34.50, 38.55] | 68.17 [66.32, 70.01] | 41.30 [39.28, 43.35] |
| *Temporal (*L-T*)* | | | | | | |
| NS-DPO | 75.74 [73.98, 77.45] | 46.06 [44.03, 48.09] | 63.87 [62.11, 65.62] | 36.26 [34.19, 38.33] | 68.06 [66.27, 69.89] | 41.02 [38.99, 43.14] |
| *Error-layer (*L-$\varepsilon$*)* | | | | | | |
| **H-DPO (ours)** | 76.82 [75.09, 78.48] | 48.11 [46.05, 50.19] | 65.04 [63.34, 66.70] | 37.55 [35.52, 39.63] | 68.06 [66.25, 69.92] | 41.02 [38.98, 43.10] |

Absolute win rates are therefore not directly comparable to published leaderboard numbers; the H-DPO-versus-DPO paired contrasts, which use the same judge, prompts, and decoding for both methods, remain internally valid.

RewardBench uses the standard DPO implicit reward

$$\hat{r}(x, y) = \beta \sum_t \big[\log \pi(y_t \mid x, y_{<t}) - \log \pi_{\text{ref}}(y_t \mid x, y_{<t})\big]$$

summed over answer tokens. SimPO is the only exception: following its official reference-free convention, we score it by average per-token log-probability rather than by a policy-to-reference log-ratio. Its RewardBench entries are therefore reported for completeness but are not directly commensurable with the log-ratio rows.

For open-ended generation, all methods are decoded under the same benchmark-specific settings for a given base model. The H-DPO-versus-DPO paired comparisons in the main text are always computed on the same prompts, with the same judge and decoding protocol.

### G.6 FULL DISCRIMINATION RESULTS

Table 9 reports RewardBench overall scores with 95% bootstrap intervals over benchmark prompts. These are the most direct real-world measurements of the preference-discrimination axis emphasized in the main text.

### G.7 FULL GENERATION RESULTS

Tables 10–12 report the open-ended generation metrics for each base model. Following the main text, we treat these results as descriptive rather than as the primary empirical claim. Within each base model, all methods are evaluated under the same decoding and judging protocol.

Table 10: Open-ended generation results on Llama-3.1-8B-Instruct. AE2 denotes AlpacaEval 2.0 raw win rate (WR) and length-controlled win rate (LC); WR is reported with a 95% bootstrap interval, and LC with its analytic standard error from the official length-control GLM. AH denotes Arena-Hard headline Elo win rate (AH Elo), per-prompt raw score (AH raw), and style-controlled win rate (AH SC). AH Elo is reported without an interval because bootstrapping it would require refitting the internal MLE.

| Method | AE2 WR | AE2 LC | AH Elo | AH raw | AH SC |
|---|---|---|---|---|---|
| *Homoscedastic references* | | | | | |
| DPO | 26.71 [23.73, 29.81] | 23.61 ±1.00 | 38.43 | 43.76 [40.93, 46.64] | 38.10 [33.44, 42.81] |
| IPO | 23.60 [20.75, 26.71] | 20.83 ±1.14 | 37.69 | 43.16 [40.28, 46.07] | 37.80 [33.32, 42.57] |
| SimPO | 23.35 [20.37, 26.34] | 22.41 ±0.89 | 36.57 | 42.54 [39.65, 45.49] | 35.97 [30.94, 41.31] |
| *Loss-layer (*L-X*)* | | | | | |
| $\beta$-DPO | 20.87 [18.14, 23.73] | 19.29 ±0.96 | 34.73 | 41.97 [39.11, 44.88] | 35.19 [30.88, 39.68] |
| $\varepsilon$-DPO | 23.98 [21.11, 26.96] | 20.88 ±0.69 | 38.96 | 44.30 [41.32, 47.26] | 39.09 [34.02, 44.29] |
| *Utility-side (*L-V*)* | | | | | |
| MixDPO | 21.49 [18.63, 24.22] | 18.71 ±0.99 | 34.65 | 41.73 [38.95, 44.60] | 34.95 [30.62, 39.66] |
| EM-DPO | 25.22 [22.24, 28.20] | 22.41 ±0.90 | 37.75 | 42.98 [40.00, 45.95] | 37.83 [32.70, 43.22] |
| *Temporal (*L-T*)* | | | | | |
| NS-DPO | 24.22 [21.24, 27.20] | 21.99 ±0.85 | 39.17 | 44.24 [41.41, 47.14] | 39.23 [34.49, 44.65] |
| *Error-layer (*L-$\varepsilon$*)* | | | | | |
| **H-DPO (ours)** | 27.95 [24.84, 31.06] | 24.39 ±0.96 | 40.48 | 44.76 [41.86, 47.77] | 40.57 [35.89, 45.37] |

Table 11: Open-ended generation results on Gemma-2-9B-it. AE2 denotes AlpacaEval 2.0 raw win rate (WR) and length-controlled win rate (LC); WR is reported with a 95% bootstrap interval, and LC with its analytic standard error from the official length-control GLM. AH denotes Arena-Hard headline Elo win rate (AH Elo), per-prompt raw score (AH raw), and style-controlled win rate (AH SC). AH Elo is reported without an interval because bootstrapping it would require refitting the internal MLE.

| Method | AE2 WR | AE2 LC | AH Elo | AH raw | AH SC |
|---|---|---|---|---|---|
| *Homoscedastic references* | | | | | |
| DPO | 48.07 [44.72, 51.43] | 47.62 ±0.67 | 57.97 | 55.56 [52.56, 58.39] | 58.94 [54.03, 63.99] |
| IPO | 53.66 [50.19, 57.15] | 51.92 ±0.69 | 59.85 | 56.65 [53.64, 59.64] | 59.88 [55.12, 64.69] |
| SimPO | 51.43 [48.07, 54.78] | 52.35 ±0.65 | 59.75 | 56.66 [53.53, 59.72] | 59.86 [55.23, 64.75] |
| *Loss-layer (*L-X*)* | | | | | |
| $\beta$-DPO | 43.23 [39.88, 46.71] | 44.69 ±0.75 | 52.02 | 52.12 [49.15, 55.07] | 52.65 [47.67, 57.83] |
| $\varepsilon$-DPO | 55.53 [52.05, 59.01] | 54.74 ±0.79 | 56.48 | 54.70 [51.80, 57.70] | 56.91 [51.93, 62.12] |
| *Utility-side (*L-V*)* | | | | | |
| MixDPO | 39.38 [36.02, 42.86] | 41.37 ±0.75 | 53.19 | 52.91 [49.97, 55.84] | 54.42 [49.50, 59.72] |
| EM-DPO | 48.45 [44.97, 51.93] | 48.26 ±0.68 | 59.90 | 56.49 [53.58, 59.39] | 61.18 [56.26, 66.23] |
| *Temporal (*L-T*)* | | | | | |
| NS-DPO | 50.56 [47.08, 54.04] | 50.98 ±0.72 | 60.44 | 57.14 [54.06, 60.15] | 61.57 [56.50, 66.62] |
| *Error-layer (*L-$\varepsilon$*)* | | | | | |
| **H-DPO (ours)** | 48.32 [44.97, 51.80] | 48.79 ±0.71 | 60.28 | 57.41 [54.36, 60.37] | 62.06 [56.93, 67.33] |

### G.8 Paired H-DPO-versus-DPO comparisons

Table 13 reports paired stratified bootstrap comparisons between H-DPO and DPO on the same base model, judge, and prompts. Positive $\Delta$ values favor H-DPO. These paired comparisons, rather than overlap or non-overlap of marginal intervals, are the basis for all method-versus-method statements in the paper.

**How to read the intervals.** The per-method intervals in Table 9 and in Tables 10–12 are *marginal*: they quantify uncertainty for one method at a time. Overlap of marginal intervals therefore does *not* imply that two methods are indistinguishable, and we do not use them that way. All method-versus-method claims in the paper rest on the paired comparisons in Table 13, which resample the shared prompts and therefore remove prompt-level variance common to both arms.

Table 12: Open-ended generation results on Mistral-7B-Instruct-v0.3. AE2 denotes AlpacaEval 2.0 raw win rate (WR) and length-controlled win rate (LC); WR is reported with a 95% bootstrap interval, and LC with its analytic standard error from the official length-control GLM. AH denotes Arena-Hard headline Elo win rate (AH Elo), per-prompt raw score (AH raw), and style-controlled win rate (AH SC). AH Elo is reported without an interval because bootstrapping it would require refitting the internal MLE.

| Method | AE2 WR | AE2 LC | AH Elo | AH raw | AH SC |
|---|---|---|---|---|---|
| *Homoscedastic references* | | | | | |
| DPO | 19.63 [16.89, 22.36] | 15.45 ±0.64 | 23.66 | 32.57 [29.73, 35.42] | 23.91 [20.37, 27.71] |
| IPO | 18.14 [15.53, 20.87] | 16.20 ±0.63 | 22.82 | 32.64 [29.83, 35.45] | 21.74 [17.99, 25.38] |
| SimPO | 19.50 [16.77, 22.24] | 17.65 ±0.74 | 19.72 | 28.82 [26.02, 31.77] | 19.23 [15.85, 22.85] |
| *Loss-layer (*L-X*)* | | | | | |
| $\beta$-DPO | 14.41 [12.05, 16.89] | 13.92 ±0.71 | 19.61 | 30.93 [28.35, 33.60] | 20.01 [17.03, 23.30] |
| $\varepsilon$-DPO | 18.88 [16.15, 21.49] | 12.32 ±0.51 | 21.14 | 30.31 [27.48, 33.12] | 21.35 [17.98, 24.90] |
| *Utility-side (*L-V*)* | | | | | |
| MixDPO | 10.06 [8.07, 12.17] | 10.75 ±0.75 | 16.63 | 28.81 [26.40, 31.36] | 17.13 [14.33, 20.12] |
| EM-DPO | 19.50 [16.77, 22.24] | 15.57 ±0.60 | 23.20 | 31.83 [28.99, 34.69] | 23.45 [19.81, 27.19] |
| *Temporal (*L-T*)* | | | | | |
| NS-DPO | 19.88 [17.14, 22.61] | 15.63 ±0.66 | 23.57 | 31.94 [29.13, 34.84] | 23.30 [19.49, 27.11] |
| *Error-layer (*L-$\varepsilon$*)* | | | | | |
| **H-DPO (ours)** | 22.48 [19.63, 25.34] | 17.74 ±0.57 | 24.81 | 33.15 [30.23, 36.05] | 24.67 [20.71, 28.61] |

Table 13: Paired stratified bootstrap ($B = 10^4$) for H-DPO versus DPO on the same base model, judge, and prompts. Positive $\Delta$ favors H-DPO. Length-controlled rows report the analytic $z$ statistic of the GLM contrast instead of a bootstrap interval.

| Base | Metric | $\Delta$ (pp) | 95% CI | Verdict |
|---|---|---|---|---|
| Llama-3.1-8B | RB v1 | +1.06 | [+0.13, +1.99] | significant |
| Llama-3.1-8B | RB v2 | +1.78 | [+0.56, +2.96] | significant |
| Llama-3.1-8B | AE2 WR | +1.24 | [−1.37, +3.85] | n.s. |
| Llama-3.1-8B | AE2 LC | +0.78 | $z$=+1.10 | n.s. |
| Llama-3.1-8B | AH raw | +1.02 | [−1.30, +3.33] | n.s. |
| Gemma-2-9B | RB v1 | +1.44 | [+0.67, +2.19] | significant |
| Gemma-2-9B | RB v2 | +1.36 | [+0.55, +2.21] | significant |
| Gemma-2-9B | AE2 WR | +0.25 | [−2.61, +3.11] | n.s. |
| Gemma-2-9B | AE2 LC | +1.17 | $z$=+2.35 | significant |
| Gemma-2-9B | AH raw | +1.85 | [−0.38, +4.03] | n.s. |
| Mistral-7B | RB v1 | +0.05 | [−0.42, +0.52] | n.s. |
| Mistral-7B | RB v2 | +0.16 | [−0.60, +0.89] | n.s. |
| Mistral-7B | AE2 WR | +2.86 | [+0.12, +5.59] | significant |
| Mistral-7B | AE2 LC | +2.29 | $z$=+5.23 | significant |
| Mistral-7B | AH raw | +0.62 | [−1.45, +2.70] | n.s. |

**Multiplicity.** Table 13 reports 15 uncorrected paired comparisons; the Bonferroni calculation below is included only as a conservative sensitivity check. Under a Bonferroni correction at $\alpha = 0.05/15 \approx 0.0033$, the surviving effects are the two RewardBench contrasts on Gemma-2-9B-it and the length-controlled contrast on Mistral-7B-Instruct-v0.3. We therefore take the preference-discrimination effect as the primary empirical claim from these real-world comparisons, and treat the remaining generation-side contrasts as descriptive.

### G.9 Reproducibility and availability

Every real-world configuration needed to reproduce our results is documented in this appendix section: the training corpus and its per-base held-out splits, the shared training configuration and the base-specific differences, the per-method hyperparameter grids together with the per-base $\lambda_\tau$ grids and selection described above, and the evaluation protocol, including the judge, the decoding settings, and the RewardBench scoring rule. All DPO-family runs use the public `swift rlhf` interface with the `loss_type` settings listed above, so the documented protocol fully determines each

run. The synthetic experiment is likewise fully specified by its appendix protocol, including the generative model and the global validation-NLL $\lambda_\tau$-selection routine (Appendix F). Anonymized code and configuration files sufficient to reproduce all reported runs are provided as supplementary material.

# H Real-World Evidence for the Learned Comparison Scale

This appendix provides additional real-world evidence that the learned H-DPO scale behaves like the comparison-specific quantity posited by the error-layer interpretation. Since no ground-truth latent scale is observable in real data, our goal is not latent-scale recovery but behavioral and statistical validation: we test whether the learned scale aligns with an external ambiguity proxy, varies materially across response pairs under the same prompt, and predicts a behavioral correlate under an external judge. We then provide qualitative examples and briefly relate these findings to the mechanism controls in Section 5.3.

## H.1 Experimental setting

Unless otherwise stated, all analyses in this appendix use the archived outputs of the Llama-3.1-8B-Instruct H-DPO run with $\beta = 0.01$ and $\lambda_\tau = 0.1$, corresponding to the main-text configuration denoted `dpo_hdpo_b0.01_tau0.1`. We focus on this run because it is the main real-world ablation base in Section 5.3 and provides the cleanest like-for-like setting for the additional diagnostics reported here.

The analyses use three benchmark sources:

- **RewardBench v1**, for the external ambiguity-proxy analysis;
- **RewardBench v2**, for the within-prompt variance decomposition; and
- **Arena-Hard**, for the external-judge positional-instability analysis.

All reported statistics are computed from archived JSON outputs.

## H.2 Alignment with an external ambiguity proxy

We first test whether the learned comparison scale $\tau_\phi$ tracks semantic ambiguity on RewardBench v1. The analysis uses 2,985 response pairs. As an external ambiguity proxy, we compute the cosine similarity between BGE-large embeddings (Xiao et al., 2024) of the two candidate responses. This proxy is produced by an external retrieval encoder that is disjoint from the policy model, so it is not literally one of the scale head's inputs. We stress, however, that it is *construct-analogous* to one of the head's seven scalar features, namely the cosine similarity between the two response representations (Appendix E): both quantities measure how close the two responses are in an embedding space and differ only in which encoder defines that space. The correlation below should therefore be read as confirming that the head's learned scale agrees with a response-similarity signal defined in an independent embedding space, not as evidence about a construct the head never observes.

The intuition is that, all else equal, semantically similar responses are often harder to distinguish than semantically dissimilar ones. Semantic similarity is therefore not a complete definition of ambiguity, but it is a natural external proxy for one important source of pairwise comparison difficulty.

**Zero-order association.** The zero-order Spearman association between learned scale and semantic similarity is strong and positive:

$$\rho_{\mathrm{S}}(\tau_\phi,\ \text{semantic similarity}) = 0.589, \qquad p < 0.001. \tag{74}$$

Thus, pairs whose responses are more semantically similar tend to receive larger learned scales.

**Controlling for natural confounds.** A natural concern is that this relationship might be explained by simpler factors, such as response length, benchmark subset, or an external reward margin. To test this, we compute a partial rank association controlling for:

Table 14: Within-prompt variation of the learned scale on RewardBench v2 subsets. Within-prompt variation is present in every subset, although its magnitude varies substantially across subsets.

| Subset | Within-prompt share | ICC | Prompts |
|---|---|---|---|
| Math | 0.660 | 0.340 | 183 |
| Safety | 0.417 | 0.583 | 450 |
| Focus | 0.395 | 0.605 | 495 |
| Factuality | 0.286 | 0.714 | 475 |
| Precise IF | 0.152 | 0.848 | 160 |

- an external reward margin,
- chosen-response length,
- rejected-response length,
- absolute length difference, and
- RewardBench subset.

Here the external reward margin is computed from the archived benchmark-side reward signal rather than from the learned scale itself. Operationally, we residualize both variables with respect to the control set and then compute Spearman correlation between the residuals; we report this as a partial rank association. Confidence intervals are obtained by bootstrap over pairs. After controlling for these variables, the association remains significant:

$$\rho_{\mathrm{S,partial}} = 0.298, \qquad 95\%\ \mathrm{CI} = [0.265, 0.332], \qquad p < 0.001. \tag{75}$$

This result shows that the learned scale is not merely a task-category or response-length artifact. Even after accounting for several natural confounds, closely matched responses still receive larger learned scales, as predicted by the observation-scale interpretation. This should be read as evidence of alignment with an external ambiguity indicator, not as proof that semantic similarity is the only or definitive source of comparison ambiguity.

### H.3 Substantial within-prompt variation in the learned scale

We next test whether the learned scale can be explained by prompt difficulty alone, or whether it varies materially across response pairs under the same prompt. RewardBench v2 is useful for this purpose because it contains multiple response comparisons for the same prompt. We use the five non-Ties subsets, comprising 1,763 prompts and 5,289 response pairs.

We work in log-scale space because the model parameterizes scale multiplicatively. Treating prompts as groups, we decompose the variance of $\log \tau_\phi$ into between-prompt and within-prompt components using a one-way random-effects decomposition. The reported intraclass correlation coefficient (ICC) is the between-prompt share of total variance, so the within-prompt share is $1 - \mathrm{ICC}$.

**Variance decomposition.** The decomposition yields

$$\sigma^2_{\mathrm{within}} = 0.005959, \qquad \sigma^2_{\mathrm{between}} = 0.004734. \tag{76}$$

The within-prompt variance share is therefore

$$\frac{\sigma^2_{\mathrm{within}}}{\sigma^2_{\mathrm{within}} + \sigma^2_{\mathrm{between}}} = 0.557, \qquad \mathrm{ICC} = 0.443. \tag{77}$$

Thus, more than half of the observed log-scale variation occurs across different response pairs under the same prompt. This rules out a purely prompt-level interpretation of the learned scale and shows that prompt difficulty alone is insufficient to explain it.

Table 15: Arena-Hard positional self-disagreement by learned-scale tercile. Larger learned scales are associated with greater positional instability under the same external judge.

| Learned-scale tercile | Mean disagreement |
|---|---|
| Low | 0.092 |
| Medium | 0.132 |
| High | 0.135 |

**Subset-level decomposition.** Table 14 shows the same decomposition within individual RewardBench v2 subsets.

Within-prompt variation is present in every subset, supporting the interpretation of $\tau_\phi$ as a comparison-specific quantity. At the same time, the magnitude of this effect varies across subsets, which is consistent with the idea that some benchmark categories contain more pair-specific ambiguity than others. This analysis does not imply that prompt difficulty is irrelevant; rather, it shows that prompt-level variation is not sufficient to explain the learned scale.

### H.4 A BEHAVIORAL CORRELATE UNDER AN EXTERNAL JUDGE

We next ask whether the learned scale predicts a behavioral signature of difficult comparisons under an external judge. Arena-Hard evaluates each model–baseline pair twice with answer positions swapped, which provides a natural test of positional stability.

For 496 valid model–baseline pairs, we define *positional self-disagreement* as the absolute difference between the two position-normalized judge scores obtained under the two answer orders, using the benchmark's standard position-normalized archived outputs. If larger learned scales correspond to harder or more ambiguous comparisons, they should be associated with larger positional disagreement under the same judge.

We find a positive Spearman association:

$$\rho_{\mathrm{S}}(\tau_\phi,\ \text{position disagreement}) = 0.129, \qquad 95\%\ \mathrm{CI} = [0.042, 0.216], \qquad p = 0.004. \tag{78}$$

Table 15 shows the same pattern when pairs are grouped by learned-scale terciles.

This result provides a deliberately limited external behavioral correlate: pairs assigned larger learned scales are more sensitive to answer position under the same judge, which is consistent with greater comparison ambiguity. Positional self-disagreement is not the same as inter-judge disagreement, and the observed effect size is modest.

### H.5 QUALITATIVE EXAMPLES

Qualitative inspection of high-scale and low-scale pairs also matches the intended semantics. On RewardBench v2, the highest-scale pairs tend to involve closely matched long-form responses, such as alternative versions of the same screenplay, code implementation variants, and similarly formatted factual lists. By contrast, the lowest-scale pairs are dominated by objectively checkable mathematical questions in which one solution is correct and the other contains a clear error.

The examples in Table 16 are drawn from the extreme tails of the learned-scale distribution and summarized in abbreviated form for readability. The absolute $\tau_\phi$ values are specific to this trained run and should be interpreted comparatively within-run rather than as calibrated scale units across runs.

These examples are illustrative rather than exhaustive, but they provide an intuitive sanity check that agrees with the quantitative analyses above.

### H.6 RELATION TO MECHANISM CONTROLS

The analyses above address the *semantics* of the learned scale: whether it behaves like a meaningful comparison-specific quantity in real data. A separate question is whether the empirical gains of

Table 16: Representative qualitative examples of learned comparison scale on RewardBench v2. Large scales correspond to subtle, closely matched comparisons; small scales correspond to sharply discriminable pairs.

| $\tau_\phi$ | Subset | Pair characteristic |
|---|---|---|
| 1.151 | Precise IF | Two nearly identical versions of the same screenplay |
| 1.144 | Precise IF | Closely matched long-form response variants |
| 1.139 | Factuality | Two similarly formatted animated-film lists |
| 0.076 | Math | Standard geometry solution versus incorrect reasoning |
| 0.059 | Math | Correct versus incorrect joint-density derivation |
| 0.014 | Math | Correct versus incorrect compound-return calculation |

H-DPO could be explained simply by adding an auxiliary head or generic pair-adaptive capacity. Section 5.3 addresses that question through capacity-matched controls; we summarize the connection here briefly.

In particular, the learned weighted-DPO control, which uses the same pair features and an almost identical head but applies the output as a loss weight rather than an inside-logit scale, is substantially worse than full H-DPO on RewardBench v1 and v2 by 5.82 and 4.58 points, respectively. A prompt-only scale parameterization is also weaker than the full model, which, together with the within-prompt variance decomposition above, argues against a prompt-only explanation of the learned scale. This decomposition is descriptive rather than causal: it shows that prompt identity alone does not account for the learned scale, but it does not by itself identify the source of the remaining within-prompt variation.

Taken together, these controls support the conclusion that the gains of H-DPO are not explained by an auxiliary head alone or by generic loss-level pair adaptation. The inside-logit comparison-scale mechanism is important.

### H.7 Scope and takeaway

Taken together, these analyses support a focused claim: the learned H-DPO scale behaves like a meaningful comparison-specific quantity in real data, consistent with the observation-scale interpretation of the error-layer branch. It aligns with an external semantic-ambiguity proxy, exhibits substantial within-prompt variation across response pairs, and predicts a modest but significant behavioral correlate under an external judge. Combined with the capacity-matched controls in Section 5.3, these results argue against interpreting the scale as a prompt-only artifact or as a generic auxiliary degree of freedom. No single analysis is decisive on its own; the point is the consistency of the pattern across several complementary diagnostics.

At the same time, these analyses do not establish recovery of a ground-truth latent scale, do not imply that all real preference heterogeneity is error-layer heterogeneity, and do not establish that H-DPO significantly improves every downstream generation metric. Their role is narrower: to show that the learned scale exhibits the qualitative and statistical signatures expected of comparison-specific observation scale in real binary preference data.

## I Ablation Details

This appendix gives the full protocol for the ablations summarized in Section 5.3. The ablation suite is designed to isolate the source of the empirical effect observed for H-DPO on the discrimination axis. In particular, it asks whether the gain comes from the heteroscedastic inside-logit mechanism itself or merely from adding an auxiliary head, extra features, or pair-adaptive weighting.

### I.1 Ablation suite

All ablations are run on **Llama-3.1-8B-Instruct** only, using the same training data, reference model, judge, decoding configuration, and attention implementation as the corresponding main-table run. Unless explicitly changed by the ablation definition, all settings are inherited from the main Llama

Table 17: Ablation suite with marginal 95% intervals on Llama-3.1-8B-Instruct. AE2 LC reports the analytic standard error of the official length-control GLM. Arena-Hard headline Elo is omitted here because its internal MLE would require refitting for interval estimation; the main text reports the corresponding point estimate. Intervals are marginal, so overlap does not license an indistinguishability claim; paired comparisons are reported separately in Table 18.

| Variant | AE2 LC | AE2 WR | AH SC | RB v1 | RB v2 |
|---|---|---|---|---|---|
| DPO (reference) | 23.61 ±1.00 | 26.71 [23.73, 29.81] | 38.10 [33.43, 42.96] | 75.76 [74.07, 77.47] | 46.32 [44.27, 48.37] |
| H-DPO (full) | 24.39 ±0.96 | 27.95 [24.97, 31.18] | 40.57 [35.87, 45.29] | 76.82 [75.08, 78.56] | 48.11 [46.08, 50.16] |
| *$\tau$-division removed* | | | | | |
| No division: additive head ($\tau = 1$) | 23.12 ±0.88 | 25.59 [22.61, 28.70] | 38.24 [33.47, 43.17] | 75.57 [73.80, 77.26] | 47.69 [45.58, 49.74] |
| No division: loss weight (W-DPO) | 21.23 ±0.88 | 22.48 [19.63, 25.47] | 38.54 [33.95, 43.42] | 71.00 [69.16, 72.77] | 43.52 [41.50, 45.60] |
| *$\tau$-division retained, scale function degraded* | | | | | |
| With division: no frozen-reference cues | 22.61 ±0.86 | 25.34 [22.36, 28.45] | 37.81 [33.03, 42.60] | 75.56 [73.84, 77.27] | 47.88 [45.81, 49.91] |
| With division: no prompt features | 22.55 ±0.95 | 25.34 [22.36, 28.45] | 40.75 [35.71, 46.00] | 76.36 [74.62, 78.09] | 47.73 [45.62, 49.77] |
| With division: prompt features only | 21.06 ±0.99 | 24.10 [21.24, 27.08] | 36.80 [31.80, 41.78] | 75.65 [73.89, 77.39] | 48.00 [45.97, 50.09] |

H-DPO configuration, including the selected scale regularization coefficient $\lambda_\tau = 0.1$. The purpose is to make each row a like-for-like comparison against the full H-DPO objective within the same base-model setting.

**Variant definitions.** The ablation variants fall into two groups. The first group tests whether the inside-logit mechanism itself matters once auxiliary capacity is held roughly fixed; the second tests whether the effect depends on particular feature families once the heteroscedastic objective is retained.

**Removing the inside-logit $\tau$-division.** These variants keep the auxiliary head but remove the defining mechanism of H-DPO.

- **No division: additive head ($\tau = 1$).** The scale head is retained with the same feature set and parameter count, but its scalar output is added to the preference logit rather than dividing it, and $\tau$ is fixed to one. This is the capacity-matched control for the inside-logit division structure.
- **No division: loss weight (W-DPO).** The same feature pipeline drives a learnable per-pair weight $w_i$ multiplying the per-pair loss term, while the comparison model itself remains homoscedastic. This is the real-world counterpart of the oracle weighted-DPO control in Section 5.1.

**Retaining $\tau$-division while degrading the scale function.** These variants keep the H-DPO objective but weaken the information available to the scale head.

- **With division: no frozen-reference cues.** Frozen-reference-derived ambiguity cues are removed from the scale-head input.
- **With division: no prompt features.** Prompt-conditioned features are removed from the scale-head input.
- **With division: prompt features only.** Only prompt-conditioned features are retained.

The feature-mask variants are implemented by zeroing the corresponding feature blocks before the head, so the head architecture and parameter count remain unchanged.

The ablation suite supports two conclusions. First, the evidence points to the inside-logit $\tau$-division as the main mechanism behind the discrimination gain. Replacing it with an additive control weakens the effect and does not match the full objective, while routing the same information to a loss-layer weight is substantially worse. This is the real-world counterpart of the synthetic weighted-DPO control in Section 5.1: pair-adaptive weighting does not reproduce what changing the comparison channel itself achieves.

Second, the gain is not reducible to any single feature family alone. On RewardBench v2 the feature-mask degradations ($-0.11$ to $-0.38$) are smaller than either mechanism-removal control (additive head $-0.42$, loss weight $-4.58$). On RewardBench v1 the picture is more mixed: removing the

Table 18: Paired stratified bootstrap ($B = 10^4$) against the full H-DPO objective on the same prompts, judge, and base model. Negative $\Delta$ favors the full objective. We report only the two mechanism-level controls most directly tied to the inside-logit-versus-loss-layer question in Section 5.3.

| Variant | Metric | $\Delta$ (pp) | 95% CI | Verdict |
|---|---|---|---|---|
| No division: additive head ($\tau = 1$) | RB v1 | $-1.25$ | $[-2.31, -0.18]$ | significant |
| No division: additive head ($\tau = 1$) | RB v2 | $-0.42$ | $[-1.63, +0.79]$ | n.s. |
| No division: additive head ($\tau = 1$) | AE2 WR | $-2.36$ | $[-5.18, +0.46]$ | n.s. |
| No division: loss weight (W-DPO) | RB v1 | $-5.82$ | $[-7.19, -4.44]$ | significant |
| No division: loss weight (W-DPO) | RB v2 | $-4.58$ | $[-6.20, -2.95]$ | significant |
| No division: loss weight (W-DPO) | AE2 WR | $-5.47$ | $[-8.20, -2.73]$ | significant |

Table 19: Scale-regularization sweep on Llama-3.1-8B-Instruct with marginal 95% intervals. $\lambda_\tau$ is the coefficient of Eq. (10); $\beta = 0.01$ is fixed. AE2 LC reports the analytic standard error of the official length-control GLM; Arena-Hard style-controlled (SC) and RewardBench intervals are marginal bootstraps over benchmark prompts ($B = 10^4$). The main table uses $\lambda_\tau = 0.1$.

| $\lambda_\tau$ | AE2 LC | AE2 WR | AH SC | RB v1 | RB v2 |
|---|---|---|---|---|---|
| 0 | 23.23 $\pm$0.97 | 25.59 [22.61, 28.70] | 40.12 [35.25, 45.19] | 75.05 [73.32, 76.76] | 48.16 [46.08, 50.27] |
| $10^{-4}$ | 23.71 $\pm$1.02 | 26.34 [23.35, 29.44] | 36.64 [31.91, 41.49] | 74.82 [73.08, 76.51] | 47.35 [45.27, 49.40] |
| $5 \times 10^{-4}$ | 23.02 $\pm$0.94 | 25.09 [22.11, 28.07] | 39.57 [35.08, 44.28] | 75.04 [73.33, 76.77] | 46.90 [44.82, 48.97] |
| $10^{-3}$ | 21.90 $\pm$0.84 | 24.22 [21.24, 27.20] | 38.82 [33.89, 43.99] | 75.54 [73.83, 77.29] | 47.43 [45.31, 49.47] |
| $5 \times 10^{-3}$ | 23.11 $\pm$1.03 | 25.59 [22.61, 28.57] | 41.58 [36.75, 46.57] | 75.93 [74.22, 77.63] | 47.64 [45.60, 49.74] |
| $10^{-2}$ | 24.57 $\pm$1.03 | 27.33 [24.22, 30.43] | 39.71 [34.90, 44.59] | 75.23 [73.49, 76.96] | 47.56 [45.58, 49.67] |
| $5 \times 10^{-2}$ | 24.32 $\pm$0.97 | 27.33 [24.22, 30.43] | 39.00 [34.09, 44.05] | 75.81 [74.08, 77.50] | 47.68 [45.61, 49.80] |
| $10^{-1}$ | 24.39 $\pm$0.96 | 27.95 [24.97, 31.06] | 40.57 [35.77, 45.36] | 76.82 [75.09, 78.53] | 48.11 [46.04, 50.16] |
| $5 \times 10^{-1}$ | 21.74 $\pm$0.94 | 24.35 [21.37, 27.45] | 42.30 [37.46, 47.32] | 75.57 [73.83, 77.30] | 48.35 [46.25, 50.45] |
| 1 | 23.36 $\pm$1.09 | 26.21 [23.23, 29.32] | 40.72 [35.85, 45.90] | 75.94 [74.20, 77.64] | 48.51 [46.43, 50.57] |
| DPO (ref) | 23.61 $\pm$1.00 | 26.71 [23.73, 29.69] | 38.10 [33.50, 42.84] | 75.76 [74.04, 77.46] | 46.32 [44.23, 48.41] |

frozen-reference cues ($-1.26$) is comparable to the additive-head control ($-1.25$), and only the loss-weight control is substantially worse ($-5.82$). We therefore base the mechanism claim on the paired contrast against the loss-weight control, which is significantly worse on all three reported paired comparisons (RB v1, RB v2, and AE2 WR), rather than on the additive control alone. This pattern is consistent with a mechanism-level effect from the heteroscedastic inside-logit structure rather than with a single-feature explanation.

### I.2 Sensitivity to the scale regularization coefficient

The training objective in Eq. (10) contains a single scale-regularization coefficient $\lambda_\tau$ that shrinks the local log-scale toward the homoscedastic baseline. This appendix reports a sweep of $\lambda_\tau$ on Llama-3.1-8B-Instruct to show that the value used in the main table ($\lambda_\tau = 0.1$) is not an isolated knife-edge choice within this base-model setting.

All rows fix $\beta = 0.01$, share the same reference policy, optimization schedule, decoding configuration, judge, and attention implementation as the corresponding main run, and differ only in $\lambda_\tau$. The $\lambda_\tau = 0$ row removes only the local-scale regularizer while keeping the heteroscedastic inside-logit objective itself, so it is *not* equivalent to DPO; a standard DPO run at the same $\beta$ is listed separately as the reference.

The sweep shows that the main-table choice $\lambda_\tau = 0.1$ lies in a reasonably stable region rather than at an isolated knife-edge point. RewardBench v2 remains above the DPO reference across the entire sweep, and RewardBench v1 recovers to its peak near $\lambda_\tau = 0.1$ after a dip at intermediate values. The AlpacaEval metrics degrade at both extremes ($\lambda_\tau = 0$ and $\lambda_\tau \geq 0.5$), whereas Arena-Hard style-controlled is non-monotone and attains its sweep maximum at $\lambda_\tau = 0.5$ (42.30 vs. 40.57 at $\lambda_\tau = 0.1$); we nonetheless retain $\lambda_\tau = 0.1$ because it maximizes RewardBench v1 and keeps all metrics within a stable band. Overall, the sweep is consistent with the practical role of a small but nonzero regularizer in stabilizing the learned local scale without collapsing the model back toward

the homoscedastic baseline. This sweep is intended as a robustness check for the selected Llama configuration, not as a claim that a single $\lambda_\tau$ transfers unchanged across all base-model families.