# LATENTCP-DPO: CONFORMAL PESSIMISM FOR DPO WITH FINITE-SAMPLE GUARANTEES

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

## ABSTRACT

Direct Preference Optimization (DPO) learns a response policy from human preference data, but with a limited dataset it can overfit to noisy or weakly informative preference labels. We introduce LatentCP-DPO, a framework that combines conformal prediction with pessimistic preference optimization for Bradley–Terry (BT) pairwise feedback and Plackett–Luce (PL) ranking feedback. Building on LatentCP's construction of uncertainty sets for unobserved model parameters, we address how to incorporate this uncertainty into preference optimization and characterize its effect on policy performance. Under exchangeability and a correctly specified preference model, we construct uncertainty sets for latent relative rewards with finite-sample marginal coverage from conformal prediction sets for observable preferences. We then learn a response policy by maximizing the worst-case expected preference log-likelihood over these uncertainty sets, using an independent collection of prompts and candidate responses without preference labels. Under suitable assumptions, we establish finite-sample upper bounds on regularized policy suboptimality that separate policy-class approximation error, conformal uncertainty, and statistical estimation error. For BT feedback, we further establish a worst-case lower bound for LatentCP-DPO that matches the upper bound's reciprocal dependence on the number of unlabeled optimization examples. The lower bound also shows that the uncertainty sets can induce a bias that persists as the optimization sample grows. Together, these results characterize both the statistical accuracy and potential limitations of preference optimization using conformal uncertainty sets. Experiments support the benefits of the proposed framework in finite preference datasets.

## 1 INTRODUCTION

Direct Preference Optimization (DPO) simplifies language-model alignment with human preferences by replacing separate reward-model training and reinforcement-learning optimization with a single stage of policy training (Rafailov et al., 2023). However, DPO can overfit to finite preference datasets (Gheshlaghi Azar et al., 2024; Rafailov et al., 2024), while noisy and inconsistent labels further complicate reliable preference learning (Chen et al., 2026). The standard DPO objective fits a policy-induced preference model to observed labels without explicitly quantifying statistical uncertainty in the inferred preferences. Policy updates may therefore place undue weight on comparisons that are weakly supported by the available data. This motivates incorporating statistically grounded uncertainty quantification into preference optimization.

Under a preference model, the probabilities of observed comparisons or rankings depend on unobserved reward differences between candidate responses. Although DPO avoids a separately trained reward model, its preference objective still depends on policy-induced reward differences. The uncertainty of interest therefore concerns the latent relative rewards governing the observed preferences. Conformal prediction (CP) provides a distribution-free framework for constructing prediction sets for observable outcomes with finite-sample marginal coverage under exchangeability (Shafer & Vovk, 2008; Angelopoulos & Bates, 2023). Under suitable sampling assumptions, LatentCP extends this construction to uncertainty sets for unobserved distributional parameters through a correctly specified forward model (Zheng et al., 2026). Using these sets in preference optimization raises two questions: *How should latent reward uncertainty be incorporated into the policy objective*, and *what performance guarantees can be established for the resulting policy*?



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.

To address these questions, we propose *LatentCP-DPO*, a framework that combines conformal prediction with pessimistic preference optimization for Bradley–Terry (BT) pairwise feedback and Plackett–Luce (PL) ranking feedback; see Fig. 1. Building on the latent-inference construction of Zheng et al. (2026), our contributions are as follows.

- **Preference optimization using latent reward uncertainty.** We first fit an initial (pilot) policy on a labeled pilot sample and use a separate labeled calibration sample to construct conformal prediction sets for observable preferences. Specializing the LatentCP construction to preference models, we convert these prediction sets into uncertainty sets for latent relative rewards with finite-sample marginal coverage under suitable assumptions. These sets induce a family of preference distributions. We then learn a response policy by maximizing the worst-case expected preference log-likelihood over this family, using an independent collection of prompts and candidate responses without preference labels.
- **Concrete objectives and algorithms for pairwise and ranking feedback.** For BT feedback, we derive a closed-form pessimistic objective that makes explicit how conformal uncertainty moderates policy-induced preference margins. For PL feedback, we develop a feasibility-preserving minorization–maximization (MM) algorithm to approximately solve the inner worst-case optimization problem.
- **A unified finite-sample upper bound.** Under stated assumptions, we establish an upper bound on KL-regularized policy suboptimality for the exact empirical optimizer. The bound separates policy-class approximation error, conformal uncertainty and calibration effects, and statistical estimation error from the independent unlabeled optimization sample.
- **A complementary lower bound and persistent bias.** For BT feedback, we establish a worst-case lower bound for LatentCP-DPO, conditional on suitable pilot realizations. At fixed confidence, this bound matches the upper bound's $d_{\Pi}/n_{\mathrm{opt}}$ dependence. It also shows that, with calibration sample size and coverage levels fixed, bias induced by the uncertainty sets can persist as the optimization sample grows.
- **Empirical improvements in prediction-set efficiency.** Evaluated over three matched policy seeds and 3,000 test contexts on Nectar with Qwen3-8B ($K = 6$), LatentCP-DPO reduces the average split-conformal ranking prediction-set size at a common requested coverage level by $2.78\%$ relative to Pure DPO and $10.09\%$ relative to the ROPO adaptation, while maintaining comparable pairwise accuracy ($66.6\%$ versus $66.5\%$ for Pure DPO). Experiments on Llama-3.1-8B-Instruct and Llama-2-7B-Chat show similar improvements in prediction-set efficiency. An ablation comparing the pilot and final policies further shows an improved coverage–set-size trade-off after preference optimization.

## 2 Related Work

We discuss the most relevant work here and defer additional discussion to Appendix A.

**Provable and Pessimistic Preference Optimization.** Provable preference learning has been studied under BT/PL feedback (Zhu et al., 2023b) and with general function approximation in offline RL (Zhan et al., 2024), with sharp guarantees for KL-regularized objectives and extensions to private preference labels (Zhao et al., 2025; Wu et al., 2025b). Pessimistic methods use adversarial rewards or critics (Liu et al., 2024; Ji et al., 2024), regularization and uncertainty penalties (Huang et al., 2025; Houliston et al., 2024), or policy ensembles (Barla et al., 2026). Distributionally robust preference optimization addresses noisy preference data, distribution shift, and PL ranking-label uncertainty (Wu et al., 2025a; Xu et al., 2025b; Wu et al., 2026). Closely related, Kim et al. (2025) optimize over $\chi^2$-divergence ambiguity sets of conditional preference distributions, obtaining closed-form inner solutions. Our ambiguity sets are induced by conformally calibrated uncertainty sets for latent relative rewards, and our analysis connects these sets to policy suboptimality and the size of an independent unlabeled optimization sample.

**Conformal Inference for Latent Parameters.** Related methods infer latent quantities using simulated parameter–observation pairs (Patel et al., 2024) or pseudo-outcomes for treatment effects (Alaa et al., 2023). LatentCP constructs uncertainty sets for unobserved distributional parameters without latent calibration labels, under suitable sampling assumptions and a correctly specified for-

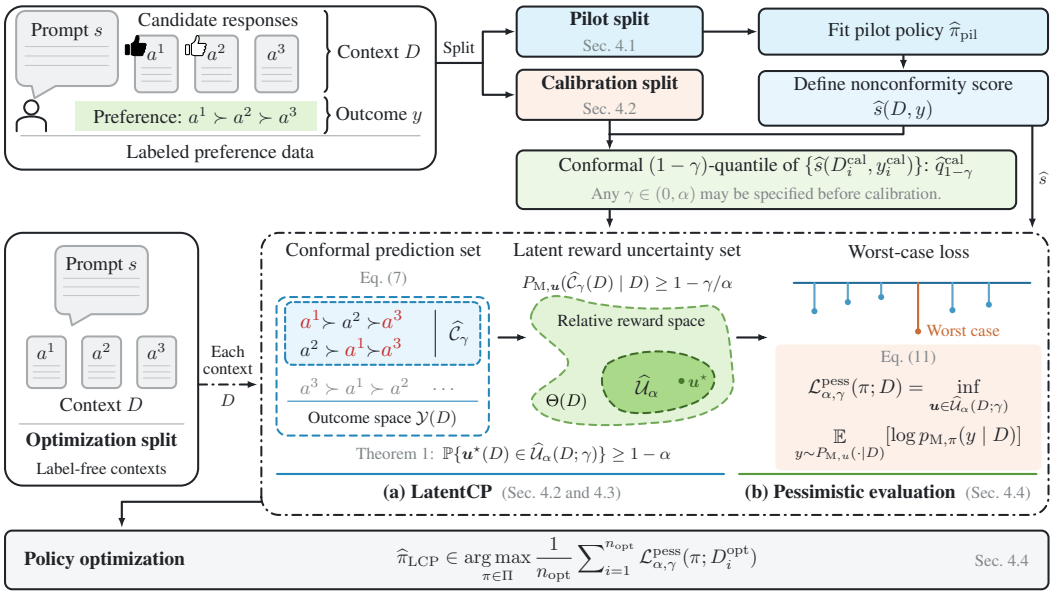


Figure 1: Overview of LatentCP-DPO framework. Conformal prediction sets for preference outcomes are inverted into uncertainty sets for latent relative rewards, which define the pessimistic policy objective. PL ranking feedback is shown for illustration.

ward model (Zheng et al., 2026). We adapt this construction to relative rewards under BT and PL feedback and establish guarantees for the resulting preference-optimization procedure.

**Conformal Alignment.** Conformal calibration has been used in robust optimization (Yeh et al., 2024) and policy risk control (Prinster et al., 2026). For alignment, conformal reward-interval widths have been used to reweight GRPO advantages (Pan et al., 2026), while response-level conformal reliability has been used to weight DPO preference pairs (Chen et al., 2026). In our framework, calibrated uncertainty sets define the preference distributions considered by the inner worst-case optimization, directly determining the pessimistic training objective.

## 3 Preliminaries

**Direct Preference Optimization (DPO).** Let $s \sim \rho$ be a prompt, let $a \in \mathcal{A}(s)$ be a response, and let $\pi_{\text{ref}}$ be a fixed reference policy. Following Rafailov et al. (2023), given a latent reward $r^\star$ and a regularization parameter $\beta > 0$, consider the KL-regularized problem

$$\pi^\star \in \arg\max_{\pi} J_\beta(\pi), \qquad J_\beta(\pi) = J(\pi) - \beta H(\pi) \tag{1}$$

where $J(\pi) = \mathbb{E}_{s\sim\rho,\, a\sim\pi(\cdot|s)}[r^\star(s,a)]$ and $H(\pi) = \mathbb{E}_{s\sim\rho}[D_{\text{KL}}(\pi(\cdot|s)\|\pi_{\text{ref}}(\cdot|s))]$. Assuming that $\pi_{\text{ref}}$ has full support and $0 < Z^\star_\beta(s) < \infty$ for $\rho$-almost every $s$, an optimizer of (1) is given by $\pi^\star(a|s) = \frac{1}{Z^\star_\beta(s)}\pi_{\text{ref}}(a|s)\exp\left(r^\star(s,a)/\beta\right)$, where $Z^\star_\beta(s) = \sum_{a'\in\mathcal{A}(s)} \pi_{\text{ref}}(a'|s)\exp(r^\star(s,a')/\beta)$. Equivalently,

$$r^\star(s,a) = \beta \log \frac{\pi^\star(a|s)}{\pi_{\text{ref}}(a|s)} + \beta \log Z^\star_\beta(s). \tag{2}$$

The final term depends on the prompt but not on the response and therefore cancels from within-prompt reward comparisons. Accordingly, for any policy $\pi$, define its implicit reward score as $r_\pi(s,a) = \beta \log \frac{\pi(a|s)}{\pi_{\text{ref}}(a|s)}$.

**Preference Models and the DPO Objective.** Let $\{(D_i, y_i)\}_{i=1}^n$ be observed preference units, where $D_i = (s_i, a_i)$ denotes the label-free context and $a_i$ is an ordered tuple of candidate responses. For pairwise feedback, $a_i = (a_i^0, a_i^1)$ and $y_i \in \{0,1\}$, where the response indices are fixed before

the preference is observed and $y_i = 1$ indicates $a_i^1 \succ a_i^0$. Under the Bradley–Terry (BT) model (Bradley & Terry, 1952), the policy-induced likelihood is

$$p_{\mathrm{BT},\pi}(y_i \mid D_i) = \sigma\big(r_\pi(s_i, a_i^1) - r_\pi(s_i, a_i^0)\big)^{y_i} \cdot \sigma\big(r_\pi(s_i, a_i^0) - r_\pi(s_i, a_i^1)\big)^{1-y_i}, \tag{3}$$

where $\sigma(t) = 1/(1 + e^{-t})$ is the logistic sigmoid function. For complete-ranking feedback, write $a_i = (a_i^1, \ldots, a_i^K)$, where $K \geq 2$, and let $y_i = (y_{i,1}, \ldots, y_{i,K})$ be a permutation of $\{1, \ldots, K\}$ that orders the candidate responses from most to least preferred. Under the Plackett–Luce (PL) model (Plackett, 1975; Luce, 2005), the policy-induced likelihood is

$$p_{\mathrm{PL},\pi}(y_i|D_i) = \prod_{k=1}^{K-1} \frac{\exp\left(r_\pi(s_i, a_i^{y_{i,k}})\right)}{\sum_{\ell=k}^{K} \exp\left(r_\pi(s_i, a_i^{y_{i,\ell}})\right)}. \tag{4}$$

Throughout, $\mathrm{M} \in \{\mathrm{BT}, \mathrm{PL}\}$ denotes the feedback model, which is fixed in advance.

Let $\Pi$ be a pre-specified policy class for which all relevant policy log-ratios and preference log-likelihoods are finite. Define the unit-level negative log-likelihood loss as $\ell_i(\pi) = -\log p_{\mathrm{M},\pi}(y_i \mid D_i)$. Assuming that the empirical minimum is attained, the empirical DPO estimator is defined by

$$\widehat{\pi}_{\mathrm{DPO}} \in \operatorname*{argmin}_{\pi \in \Pi} \frac{1}{n} \sum_{i=1}^{n} \ell_i(\pi). \tag{5}$$

**KL-Regularized Suboptimality Gap.** For a learned policy $\widehat{\pi}$ and a comparator policy $\pi^\dagger$, define the KL-regularized performance gap as $\mathrm{SubOpt}_\beta(\widehat{\pi}, \pi^\dagger) = J_\beta(\pi^\dagger) - J_\beta(\widehat{\pi})$. The comparator $\pi^\dagger$ need not be optimal, so these relative performance gaps are not necessarily nonnegative.

**Conformal Prediction.** Conformal prediction (CP), in its split form, uses a fitted predictor and a separate calibration sample to construct prediction sets for observable outcomes, with distribution-free finite-sample marginal coverage under exchangeability (Angelopoulos et al., 2024). Here, *exchangeability* means that, conditional on the data used to fit the score, the joint distribution of the calibration units and a future unit is invariant under permutations. In our setting, preference outcomes are observable, whereas relative rewards are latent; uncertainty sets for the latter are developed in Section 4.3.

## 4 The LatentCP-DPO Framework

We partition the $n$ labeled preference units into disjoint pilot and calibration subsets, $\{1, \ldots, n\} = \mathcal{I}_{\mathrm{pil}} \dot{\cup} \mathcal{I}_{\mathrm{cal}}$. We use the pilot subset to fit a pilot DPO policy and construct the nonconformity score, and the calibration subset to calibrate the conformal prediction sets. Let $D_1^{\mathrm{opt}}, \ldots, D_{n_{\mathrm{opt}}}^{\mathrm{opt}}$ be label-free preference contexts drawn from the target distribution independently of the pilot and calibration data. We use these contexts to form the sample-average pessimistic DPO objective.

### 4.1 Pilot Policy and Nonconformity Score

Let $n_{\mathrm{pil}} = |\mathcal{I}_{\mathrm{pil}}|$. Applying the empirical DPO estimator defined in (5) to the pilot subset, we obtain the pilot policy $\widehat{\pi}_{\mathrm{pil}}$. Using this pilot policy, we define the nonconformity score as

$$\widehat{s}(D, y) = -\log p_{\mathrm{M},\widehat{\pi}_{\mathrm{pil}}}(y \mid D). \tag{6}$$

A larger value of $\widehat{s}(D, y)$ corresponds to a lower probability assigned to $y$ by the pilot-induced preference distribution and, therefore, to greater nonconformity. After fitting, both $\widehat{\pi}_{\mathrm{pil}}$ and the score function $\widehat{s}$ are fixed throughout the subsequent calibration and pessimistic optimization stages.

### 4.2 Conformal Calibration for Preference Outcomes

Fix a target miscoverage level $\alpha \in (0, 1)$ and an auxiliary preference-outcome miscoverage level $\gamma \in (0, \alpha)$, chosen independently of the calibration data and the future preference unit. Let $n_{\mathrm{cal}} = |\mathcal{I}_{\mathrm{cal}}|$. For each $i \in \mathcal{I}_{\mathrm{cal}}$, define the calibration score $S_i^{\mathrm{cal}} = \widehat{s}(D_i, y_i)$ and let $S_{(1)}^{\mathrm{cal}} \leq \cdots \leq S_{(n_{\mathrm{cal}})}^{\mathrm{cal}}$ denote the corresponding order statistics. Set $k_\gamma = \lceil (n_{\mathrm{cal}} + 1)(1 - \gamma) \rceil$ and define the conformal

threshold $\widehat{q}^{\mathrm{cal}}_{1-\gamma} = S^{\mathrm{cal}}_{(k_\gamma)}$ for $k_\gamma \le n_{\mathrm{cal}}$, and $\widehat{q}^{\mathrm{cal}}_{1-\gamma} = +\infty$ for $k_\gamma = n_{\mathrm{cal}} + 1$. For any label-free context $D$, let $\mathcal{Y}(D)$ denote the set of admissible preference outcomes. The resulting conformal prediction set is

$$\widehat{\mathcal{C}}_\gamma(D) = \left\{ y \in \mathcal{Y}(D) : \widehat{s}(D, y) \le \widehat{q}^{\mathrm{cal}}_{1-\gamma} \right\}. \tag{7}$$

If the calibration preference units and a future unit $(D_{n+1}, y_{n+1})$ are exchangeable conditional on the pilot data, then $\mathbb{P}\{y_{n+1} \in \widehat{\mathcal{C}}_\gamma(D_{n+1})\} \ge 1 - \gamma$.

### 4.3 Uncertainty Sets for Latent Relative Rewards

Define the candidate-level latent reward vector[1] by $\boldsymbol{r}^\star(D) = (r^\star(s, a^1), \ldots, r^\star(s, a^K))^\top \in \mathbb{R}^K$. Because both the BT and PL models are invariant to a common additive shift applied to all candidate rewards, define the relative-reward vector $\boldsymbol{u}^\star(D)$ by centering $\boldsymbol{r}^\star(D)$. By construction, $\mathbf{1}_K^\top \boldsymbol{u}^\star(D) = 0$. Accordingly, the latent parameter space is $\Theta(D) = \{\boldsymbol{u} \in \mathbb{R}^K : \mathbf{1}_K^\top \boldsymbol{u} = 0\}$. For the fixed feedback model $\mathrm{M} \in \{\mathrm{BT}, \mathrm{PL}\}$ and any $\boldsymbol{u} \in \Theta(D)$, let $P_{\mathrm{M},\boldsymbol{u}}(\cdot \mid D)$ denote the corresponding preference distribution. Its probability mass function $p_{\mathrm{M},\boldsymbol{u}}(y \mid D)$ is obtained from (3) or (4) by replacing each candidate-level implicit reward $r_\pi(s, a^k)$ with the corresponding coordinate of $\boldsymbol{u}$.

For the $i$-th preference unit $(D_i, y_i)$, let $\boldsymbol{u}_i^\star = \boldsymbol{u}^\star(D_i) \in \Theta(D_i)$ denote its unobserved latent relative reward vector. Under the fixed feedback model M, the observable preference outcome is generated according to

$$y_i \mid (D_i, \boldsymbol{u}_i^\star) \sim P_{\mathrm{M},\boldsymbol{u}_i^\star}(\cdot \mid D_i). \tag{8}$$

Using this forward model, we define the forward $\gamma$-compatibility of any relative-reward vector $\boldsymbol{u} \in \Theta(D)$ with the conformal prediction set $\widehat{\mathcal{C}}_\gamma(D)$ as

$$p_\gamma(\boldsymbol{u}, D) = \mathbb{P}_{y \sim P_{\mathrm{M},\boldsymbol{u}}(\cdot \mid D)} \left\{ y \in \widehat{\mathcal{C}}_\gamma(D) \right\}. \tag{9}$$

Thus, $p_\gamma(\boldsymbol{u}, D)$ is the probability mass that the preference-outcome distribution induced by $\boldsymbol{u}$ assigns to $\widehat{\mathcal{C}}_\gamma(D)$. The latent relative-reward uncertainty set is defined by

$$\widehat{\mathcal{U}}_\alpha(D; \gamma) = \{ \boldsymbol{u} \in \Theta(D) : p_\gamma(\boldsymbol{u}, D) \ge 1 - \varepsilon_{\alpha,\gamma} \}, \tag{10}$$

where $\varepsilon_{\alpha,\gamma} = \gamma/\alpha$. To avoid optimizing over an empty set, whenever the set $\widehat{\mathcal{U}}_\alpha(D; \gamma)$ defined above is empty, we adopt the conservative *empty-set convention* $\widehat{\mathcal{U}}_\alpha(D; \gamma) = \Theta(D)$. We have the following coverage guarantee, with the proof provided in Appendix B.

**Theorem 1** (Finite-sample marginal coverage guarantee)**.** *Fix $\alpha \in (0, 1)$ and $\gamma \in (0, \alpha)$ independently of the calibration data and the future preference unit. Suppose that, conditional on the pilot data, the pairs $\{(D_i, \boldsymbol{u}_i^\star) : i \in \mathcal{I}_{\mathrm{cal}} \cup \{n+1\}\}$ are exchangeable. Suppose further that, conditional on both the pilot data and these pairs, the corresponding preference outcomes are independent, with respective distributions given by (8). Then $\mathbb{P}\{\boldsymbol{u}_{n+1}^\star \in \widehat{\mathcal{U}}_\alpha(D_{n+1}; \gamma)\} \ge 1 - \alpha$.*

### 4.4 From Latent Uncertainty Sets to Pessimistic DPO

For each $D$, the latent uncertainty set $\widehat{\mathcal{U}}_\alpha(D; \gamma)$ induces the ambiguity set $\{P_{\mathrm{M},\boldsymbol{u}}(\cdot \mid D) : \boldsymbol{u} \in \widehat{\mathcal{U}}_\alpha(D; \gamma)\}$, which collects the plausible conditional distributions of the preference outcome given $D$. For any candidate policy $\pi \in \Pi$, we define its unit-level pessimistic DPO objective as the worst-case expected preference log-likelihood over this ambiguity set, namely,

$$\mathcal{L}^{\mathrm{pess}}_{\alpha,\gamma}(\pi; D) = \inf_{\boldsymbol{u} \in \widehat{\mathcal{U}}_\alpha(D;\gamma)} \mathbb{E}_{y \sim P_{\mathrm{M},\boldsymbol{u}}(\cdot \mid D)} \left[ \log p_{\mathrm{M},\pi}(y \mid D) \right]. \tag{11}$$

Here, $\boldsymbol{u}$ determines the preference distribution under which the candidate policy is evaluated, while $\pi$ determines the preference log-likelihood being evaluated. We then define the LatentCP-DPO policy by maximizing the sample-average pessimistic DPO objective over the label-free optimization sample, namely,

$$\widehat{\pi}_{\mathrm{LCP}} \in \operatorname*{argmax}_{\pi \in \Pi} \frac{1}{n_{\mathrm{opt}}} \sum_{i=1}^{n_{\mathrm{opt}}} \mathcal{L}^{\mathrm{pess}}_{\alpha,\gamma}(\pi; D_i^{\mathrm{opt}}).$$

Algorithm 1 in Appendix C.1 gives a model-agnostic specification of LatentCP-DPO.

[1] Here $K$ is the number of candidate responses. For BT feedback, we use the indexing $(a^0, a^1)$, so $K = 2$ and $\boldsymbol{r}^\star(D) = (r^\star(s, a^0), r^\star(s, a^1))^\top$.

# 5 INSTANTIATIONS FOR BT AND PL MODELS

We next specialize the uncertainty sets $\widehat{\mathcal{U}}_\alpha$ and inner optimization problems (11) to the BT and PL models. The BT case admits closed-form characterizations of the latent uncertainty sets and the pessimistic objective, while the PL case is handled by a feasibility-preserving MM algorithm for approximate inner optimization.

## 5.1 BRADLEY–TERRY PAIRWISE FEEDBACK

For pairwise feedback, the latent parameter space is $\Theta(D) = \{(u^0, u^1)^\top \in \mathbb{R}^2 : u^0 + u^1 = 0\}$. Under the BT model, the preference outcome is binary, so that $\mathcal{Y}(D) = \{0, 1\}$ and $\widehat{\mathcal{C}}_\gamma(D) \in \{\varnothing, \{0\}, \{1\}, \{0, 1\}\}$. Combining (10) with the empty-set convention then yields

$$\widehat{\mathcal{U}}_\alpha(D;\gamma) = \begin{cases} \{\boldsymbol{u} \in \Theta(D) : u^1 - u^0 \geq \log \frac{\alpha-\gamma}{\gamma}\}, & \widehat{\mathcal{C}}_\gamma(D) = \{1\}, \\ \{\boldsymbol{u} \in \Theta(D) : u^1 - u^0 \leq -\log \frac{\alpha-\gamma}{\gamma}\}, & \widehat{\mathcal{C}}_\gamma(D) = \{0\}, \\ \Theta(D), & \widehat{\mathcal{C}}_\gamma(D) = \{0, 1\} \text{ or } \varnothing. \end{cases} \tag{12}$$

For any policy $\pi \in \Pi$, define the policy-induced preference margin as $z_\pi(D) = r_\pi(s, a^1) - r_\pi(s, a^0)$. Taking the infimum of the expected BT log-likelihood over this uncertainty set gives

$$\mathcal{L}^{\text{pess}}_{\alpha,\gamma}(\pi; D) = \begin{cases} \log \sigma\left(z_\pi(D)\right) - \varepsilon_{\alpha,\gamma}[z_\pi(D)]_+, & \widehat{\mathcal{C}}_\gamma(D) = \{1\}, \\ \log \sigma\left(-z_\pi(D)\right) - \varepsilon_{\alpha,\gamma}[-z_\pi(D)]_+, & \widehat{\mathcal{C}}_\gamma(D) = \{0\}, \\ \log \sigma\left(-|z_\pi(D)|\right), & \widehat{\mathcal{C}}_\gamma(D) = \{0, 1\} \text{ or } \varnothing, \end{cases} \tag{13}$$

where $[t]_+ = \max\{t, 0\}$. The detailed derivation is provided in Appendix C.2.

The BT instantiation illustrates how conformal uncertainty moderates preference optimization. For an observed label $y = 1$, the standard DPO log-likelihood $\log \sigma(z_\pi(D))$ increases as the policy-induced preference probability $\sigma(z_\pi(D))$ approaches one. In contrast, when $\widehat{\mathcal{C}}_\gamma(D) = \{1\}$, our uncertainty set allows the opposite outcome to have probability up to $\varepsilon_{\alpha,\gamma}$. The resulting one-sided penalty in (13) accounts for this possibility. For $\varepsilon_{\alpha,\gamma} < 1/2$[2], the objective, viewed as a function of an unrestricted scalar margin, is maximized when $\sigma(z_\pi(D)) = 1 - \varepsilon_{\alpha,\gamma}$: *the supported preference direction is retained, but its strength is moderated.* The case $\widehat{\mathcal{C}}_\gamma(D) = \{0\}$ is symmetric. When both outcomes are included in $\widehat{\mathcal{C}}_\gamma(D)$, or the empty-set convention applies, the unit-level objective is instead maximized at zero margin, corresponding to equal preference probabilities. These characterizations explain how the conformal uncertainty set controls the direction and strength favored by the unit-level objective. Derivations and further discussion are also provided in Appendix C.2.

## 5.2 PLACKETT–LUCE RANKING FEEDBACK

Let $\mathfrak{S}_K = \{y^{(1)}, \ldots, y^{(N_K)}\}$ denote the set of all permutations of $\{1, \ldots, K\}$, where $N_K = K!$, so that $\mathcal{Y}(D) = \mathfrak{S}_K$. The latent parameter space is $\Theta(D) = \{\boldsymbol{u} = (u^1, \ldots, u^K)^\top \in \mathbb{R}^K : \mathbf{1}_K^\top \boldsymbol{u} = 0\}$. To eliminate the zero-sum constraint, let $B_K \in \mathbb{R}^{K\times(K-1)}$ have columns forming an orthonormal basis of $\mathbf{1}_K^\perp$. Then $\Theta(D) = \{B_K \boldsymbol{v} : \boldsymbol{v} \in \mathbb{R}^{K-1}\}$.

For fixed $D$ and $\pi$, define $q_m(\boldsymbol{v}) = p_{\text{PL},B_K\boldsymbol{v}}(y^{(m)} \mid D)$, $h_m(\pi) = \log p_{\text{PL},\pi}(y^{(m)} \mid D)$, and $c_m = \mathbb{I}\{y^{(m)} \in \widehat{\mathcal{C}}_\gamma(D)\}$ for $m = 1, \ldots, N_K$. Let $\boldsymbol{q}(\boldsymbol{v})$, $\boldsymbol{h}(\pi)$, and $\boldsymbol{c}$ collect these quantities in the same enumeration order. For a nonempty conformal prediction set, the inner problem (11) admits the exact finite-dimensional reformulation

$$\mathcal{L}^{\text{pess}}_{\alpha,\gamma}(\pi; D) = \inf_{\boldsymbol{v}\in\mathbb{R}^{K-1}} \boldsymbol{q}(\boldsymbol{v})^\top \boldsymbol{h}(\pi) \quad \text{s.t.} \quad \boldsymbol{c}^\top \boldsymbol{q}(\boldsymbol{v}) \geq 1 - \varepsilon_{\alpha,\gamma}. \tag{14}$$

When $\widehat{\mathcal{C}}_\gamma(D) = \varnothing$, the compatibility constraint is omitted under the conservative empty-set convention. Here, $\boldsymbol{q}(\boldsymbol{v})$ ranges over the ranking distributions induced by the PL family. The detailed derivation is provided in Appendix C.3.

[2]Both instantiations are defined for $0 < \gamma < \alpha$. The stronger condition $\gamma < \alpha/2$ is used to obtain a nonzero unit-level optimum for a singleton BT prediction set and in the BT lower-bound construction.

Although (14) is exact, the resulting constrained problem need not be convex in $\boldsymbol{v}$ and does not generally admit a closed-form solution. We therefore develop a feasibility-preserving minorization–maximization (MM) procedure based on an equivalent maximization formulation. The derivation is provided in Appendix C.4, and the procedure is summarized in Algorithm 2. For fixed $\pi$ and $D$, let $\widetilde{\boldsymbol{v}} = \widetilde{\boldsymbol{v}}(\pi; D)$ denote the feasible approximate solution returned by the algorithm. We define the resulting approximation to the unit-level pessimistic PL objective as $\widetilde{\mathcal{L}}^{\mathrm{pess}}_{\alpha,\gamma}(\pi; D) = \boldsymbol{q}(\widetilde{\boldsymbol{v}})^\top \boldsymbol{h}(\pi)$.

In practice, we may alternate between updating each $\widetilde{\boldsymbol{v}}_i$ using the MM procedure at the current policy and updating the policy by a standard first-order method while holding the resulting ranking distributions fixed. This alternating procedure provides a practical approximation to the exact LatentCP-DPO estimator $\widehat{\pi}_{\mathrm{LCP}}$.

## 6 Theoretical Results

To establish finite-sample guarantees for the exact LatentCP-DPO estimator $\widehat{\pi}_{\mathrm{LCP}}$, we introduce the additional assumptions required for the analysis.

**Assumption 1** (Sampling for suboptimality analysis)**.** *In addition to the sampling and model-correctness conditions of Theorem 1, suppose that, conditional on the pilot sample, the calibration preference units are i.i.d. draws from the target preference-unit distribution. The label-free optimization contexts $D^{\mathrm{opt}}_1, \ldots, D^{\mathrm{opt}}_{n_{\mathrm{opt}}}$ are i.i.d. draws from its context marginal, independently of the pilot and calibration samples.*

To treat BT and PL feedback uniformly, define the policy-induced comparison score by

$$x_\pi(D) = \begin{cases} r_\pi(s, a^1) - r_\pi(s, a^0), & \mathrm{M} = \mathrm{BT}, \\ \left( r_\pi(s, a^k) - \frac{1}{K} \sum_{\ell=1}^{K} r_\pi(s, a^\ell) \right)_{k=1}^{K}, & \mathrm{M} = \mathrm{PL}, \end{cases}$$

where $K \geq 2$ is fixed across all PL preference units. Define $x^\star(D)$ analogously using $r^\star$. Thus, $x_\pi(D) = z_\pi(D)$ under BT feedback, whereas $x^\star(D) = \boldsymbol{u}^\star(D)$ under PL feedback. For BT scores, norms are interpreted as absolute values.

**Assumption 2** (Bounded Comparison Scores)**.** *There exists a constant $0 < B < \infty$ such that, for almost every target context $D$, $\sup_{\pi \in \Pi} \|x_\pi(D)\|_\infty \leq B$ and $\|x^\star(D)\|_\infty \leq B$.*

A sufficient condition is that, for $\rho$-almost every prompt $s$, $\sup_{\pi \in \Pi,\ a \in \mathcal{A}(s)} |r_\pi(s, a)| \leq B/2$ and $\sup_{a \in \mathcal{A}(s)} |r^\star(s, a)| \leq B/2$.

Define the policy-class approximation error $\mathcal{E}_\Pi = \inf_{\pi \in \Pi} \mathbb{E}_D[\|x_\pi(D) - x^\star(D)\|^2]$, which vanishes when $\pi^\star \in \Pi$. Let $\eta_\gamma = \mathbb{P}\{|\widehat{\mathcal{C}}_\gamma(D)| \neq 1\}$ denote the probability of an empty or non-singleton conformal set. We use the refined measure $\widetilde{\eta}_{\alpha,\gamma}$ defined in (45), which equals $\eta_\gamma$ for BT feedback. For PL feedback, it accounts for partial ranking information retained by non-singleton prediction sets. Here and below, expectations and probabilities over a fresh target context $D$ are conditional on the pilot and calibration samples.

For a fixed evaluation response distribution $\nu(\cdot \mid s)$, define the comparison coefficient

$$C_{\mathrm{comp}}(\nu; \Pi) = \sup_{\pi \in \Pi} \frac{\mathbb{E}_{\substack{s \sim \rho \\ a \sim \nu(\cdot|s), a' \sim \pi(\cdot|s)}} \left[ \left( [r_\pi(s, a) - r^\star(s, a)] - [r_\pi(s, a') - r^\star(s, a')] \right)^2 \right]}{\mathbb{E}_D \left[ \|x_\pi(D) - x^\star(D)\|^2 \right]} \tag{15}$$

with the convention $0/0 = 0$, where $a$ and $a'$ are conditionally independent given $s$. This coefficient measures how effectively the preference-data distribution controls reward-error differences between responses drawn from $\nu$ and a candidate policy. Larger values indicate greater error amplification and hence weaker comparison coverage. Its role is analogous to that of relative condition numbers in log-linear alignment analyses (Zhou et al., 2025).

For $\delta \in (0, 1)$, define the conformal uncertainty envelope

$$\Psi_{\alpha,\gamma}(\delta) = \min \left\{ 1,\ \varepsilon^2_{\alpha,\gamma} + (1 - \varepsilon^2_{\alpha,\gamma}) \left( \widetilde{\eta}_{\alpha,\gamma} + \gamma + \frac{\log(2/\delta)}{n_{\mathrm{cal}}} \right) \right\},$$

and the common error term

$$\mathfrak{E}_{\alpha,\gamma}(\delta) = \underbrace{\mathcal{E}_{\Pi}}_{\text{policy approximation error}} + \underbrace{\Psi_{\alpha,\gamma}(\delta)}_{\text{conformal uncertainty}} + \underbrace{\frac{d_{\Pi} + \log(2/\delta)}{n_{\text{opt}}}}_{\text{statistical estimation error}}. \tag{16}$$

Here, $d_{\Pi}$ is the complexity parameter in the localized Rademacher condition (26). Sufficient conditions for (26) are given in Lemma 2.

**Theorem 2** (Finite-Sample Suboptimality Guarantee)**.** *Suppose that Assumptions 1 and 2 and the localized-complexity condition (26) hold. Assume further that $C_{\text{comp}}(\pi^\star; \Pi) < \infty$. Fix $0 < \gamma < \alpha < 1$ independently of the calibration sample, and let $\delta \in (0, 1)$. Conditional on the pilot sample, with probability at least $1 - \delta$ over the calibration and optimization samples, every comparator policy $\pi^\dagger$ such that $J(\pi)$ and $H(\pi)$ are finite for $\pi \in \{\pi^\star, \widehat{\pi}_{\text{LCP}}, \pi^\dagger\}$ satisfies*

$$\text{SubOpt}_{\beta}(\widehat{\pi}_{\text{LCP}}, \pi^\dagger) \leq \frac{C_{\text{M}} C_{\text{comp}}(\pi^\star; \Pi)}{2\beta} \mathfrak{E}_{\alpha,\gamma}(\delta) - \beta \mathbb{E}_{s\sim\rho}\left[D_{\text{KL}}\left(\pi^\dagger(\cdot \mid s) \middle\| \pi^\star(\cdot \mid s)\right)\right], \tag{17}$$

*where $C_{\text{M}} > 0$ depends only on $B$, the model $\text{M}$, the localized-complexity constant $C_{\text{loc}}$, and, for PL feedback, the fixed number of candidates $K$. In particular, setting $\pi^\dagger = \pi^\star$ removes the comparator-dependent KL correction.*

The conformal component of the bound depends on the compatibility tolerance $\varepsilon_{\alpha,\gamma}$, the information retained by the prediction sets, and finite-sample calibration. Decreasing $\gamma$ lowers $\varepsilon_{\alpha,\gamma}$ but can enlarge the prediction sets, so its effect on the bound need not be monotone. For fixed confidence and a fixed policy class, the optimization-sample term scales as $d_{\Pi}/n_{\text{opt}}$ and decreases as more label-free optimization contexts are added. The full theorem statement and proof are provided in Appendix D. Explicit choices of $C_{\text{M}}$ are given in (27) for BT feedback and (28) for PL feedback; further discussion appears in Appendix D.1.

**Theorem 3** (Worst-case Lower Bound for LatentCP-DPO)**.** *Fix $\alpha \in (0, 1)$, $B, \beta > 0$, and $\gamma \in (0, \alpha/2)$ independently of the calibration sample. For every integer $d_{\Pi} \geq 1$ and integer sample sizes satisfying $n_{\text{pil}} \geq 2d_{\Pi} + 1$, $n_{\text{cal}} \geq 1/\gamma - 1$, and $n_{\text{opt}} \geq 8 \max\{d_{\Pi}, 1/\gamma\}$, there exist a BT instance and a log-linear policy class satisfying Assumptions 1 and 2 and the localized-complexity condition, with $\pi^\star \in \Pi$ and $C_{\text{comp}}(\pi^\star; \Pi) \leq 1$. Moreover, there exists a pilot-sample event $\mathcal{H}_{\text{pil}}$ of positive probability such that, conditional on $\mathcal{H}_{\text{pil}}$, with probability at least $1/16$ over the calibration and optimization samples, every exact LatentCP-DPO empirical optimizer satisfies*

$$\text{SubOpt}_{\beta}(\widehat{\pi}_{\text{LCP}}, \pi^\star) \geq c_{B,\beta}\left[\frac{d_{\Pi}}{n_{\text{opt}}} + (1 - 2\varepsilon_{\alpha,\gamma})^2\left(\gamma + \frac{1}{n_{\text{cal}}}\right)\right], \tag{18}$$

*where $c_{B,\beta} > 0$ depends only on $B$ and $\beta$.*

The construction and proof are provided in Appendix E. This lower bound matches the upper bound's $d_{\Pi}/n_{\text{opt}}$ dependence for the exact estimator $\widehat{\pi}_{\text{LCP}}$. It further shows that, with the calibration sample size and coverage levels fixed, conformal bias can persist as the optimization sample grows. Together, these results distinguish the estimation error reduced by additional label-free optimization contexts from the residual effects of conformal uncertainty on policy performance.

## 7 Experiments

We evaluate LatentCP-DPO under PL complete-ranking feedback, the setting with a nontrivial inner optimization problem, to assess prediction-set efficiency and pairwise preference accuracy.

**Datasets.** We use `berkeley-nest/Nectar`[3] (Zhu et al., 2023a), retaining prompts whose seven candidate responses have distinct ranks. For each prompt, we randomly permute the seven responses and keep the first $K = 6$ in permuted order, preventing candidate position from revealing the original rank. Each resulting context admits $6! = 720$ complete rankings. Using a fixed seed, we create

[3] `https://huggingface.co/datasets/berkeley-nest/Nectar`

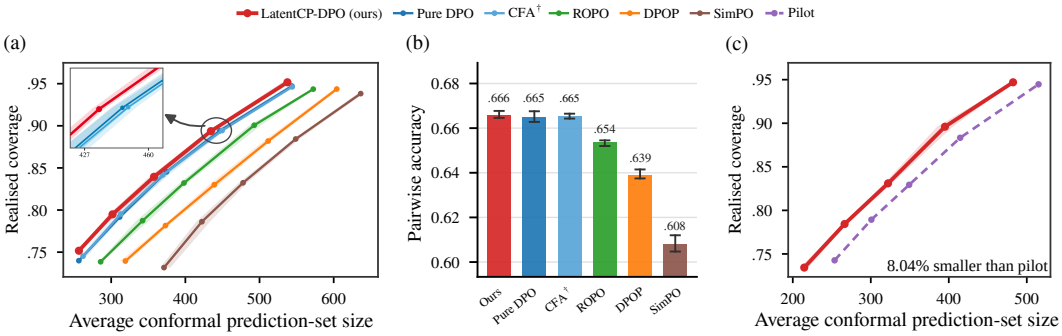


Figure 2: Results on Nectar with Qwen3-8B ($K = 6$). (a) Realized coverage versus average prediction-set size at five representative levels among the 50 evaluated nominal coverage levels. Curves show means over three matched policy seeds, and shaded bands show the seed min–max ranges. (b) Pairwise accuracy for the same policy runs and test split as in (a). Bars show seed means, and error bars show the seed ranges. (c) Descriptive coverage–set-size curves for the pilot policy and a final LatentCP-DPO policy, evaluated on different calibration and test splits from (a) and (b).

the following disjoint splits: pilot $(3,000)$, $\gamma$-calibration $(1,500)$, validation $(2,000)$, optimization $(8,000$, with $4,000$ used per final policy run), final calibration $(1,500)$, and test $(3,000)$.

**Experiment setting and compute resources.** We use Qwen3-8B for the main experiments and Llama-3.1-8B-Instruct and Llama-2-7B-Chat for the cross-architecture study in Appendix F. All policies use LoRA on attention and MLP projections ($r = 32$, $\alpha_{\text{LoRA}} = 64$, dropout $0.05$) and train for one epoch with batch size 64, AdamW (learning rate $10^{-5}$), a cosine schedule with $5\%$ warmup, gradient clipping at $1.0$, bf16, and gradient checkpointing. We fix $\beta = 0.1$, $\alpha = 0.30$, and $\gamma = 0.20$. All methods are evaluated across three matched seeds using $24\times$NVIDIA A100 GPUs.

**Metric.** For each frozen policy $\pi$, let $p_\pi(y \mid D)$ be the PL likelihood induced by its implicit rewards $\beta \log(\pi/\pi_{\text{ref}})$, and score rankings by $s(D, y) = -\log p_\pi(y \mid D)$. For 50 nominal coverage levels $\tau \in [0.75, 0.95]$, we compute the split-conformal threshold $\widehat{q}_\tau$ from that policy's scores on the shared 1,500-context final-calibration split and form $\mathcal{C}_\tau(D) = \{y \in \mathcal{Y}(D) : s(D, y) \le \widehat{q}_\tau\}$. On the $3,000$ test contexts, we report average prediction-set size against realized coverage. Each policy calibrates its own threshold at the same requested level $\tau$, so methods are compared at *matched target coverage*, the standard coverage–informativeness trade-off in conformal prediction. All 50-level results are tabulated in Appendix F.

**Results.** Figure 2 summarizes the results on Nectar with Qwen3-8B. At comparable realized coverage, LatentCP-DPO yields smaller average ranking prediction sets than all five baselines at the five displayed levels (Figure 2(a)). Across all 50 evaluated levels, the average set-size reductions at common requested coverage level are $2.78\%$, $2.82\%$, and $10.09\%$ relative to Pure DPO (Rafailov et al., 2023), the CFA adaptation (Chen et al., 2026), and the ROPO adaptation (Liang et al., 2025), respectively. Its pairwise accuracy is $66.6\%$, comparable to Pure DPO ($66.5\%$) and CFA ($66.5\%$), and numerically higher than ROPO, DPOP (Pal et al., 2024), and SimPO (Meng et al., 2024) (Figure 2(b)). The final policy also yields a smaller average set than the pilot at common requested coverage levels (Figure 2(c)).

## 8 Conclusion

We developed LatentCP-DPO, which incorporates conformally calibrated uncertainty sets for latent relative rewards into a pessimistic preference objective under BT and PL feedback. Our finite-sample upper bound on KL-regularized policy suboptimality separates policy-class approximation, conformal uncertainty, and optimization-sample estimation error, while a BT worst-case lower bound shows that conformal bias can persist as the label-free optimization sample grows. Extending this offline framework to online preference optimization remains future work.

### AI USE STATEMENT

ChatGPT (GPT-5.6 and GPT-6) was used to assist with polishing the manuscript, checking references, and refining mathematical notation. It also provided suggestions for constructing the BT instance used in the lower-bound proof. The authors reviewed all AI-assisted material and take responsibility for the final content of this work.

### ETHICS STATEMENT

This work is primarily theoretical. Our experiments evaluate offline preference optimization using the public Nectar dataset and involve no collection of new human preference judgments. We use the dataset only for research and follow its stated usage conditions. The work therefore does not constitute human-subjects research, and no IRB review was required.

### REPRODUCIBILITY STATEMENT

Section 4 introduces the LatentCP-DPO framework, and Appendix C.1 presents its model-agnostic procedure. Section 5 presents the BT and PL objectives, and Appendix C provides their derivations and the PL MM procedure. Appendices B, D, and E contain the theoretical proofs. The Experiments section reports the data selection and split sizes, training and evaluation settings, and compute resources. Appendix F provides the full 50-level results and cross-architecture comparisons.

## REFERENCES

Ahmed M Alaa, Zaid Ahmad, and Mark van der Laan. Conformal meta-learners for predictive inference of individual treatment effects. *Advances in neural information processing systems*, 36: 47682–47703, 2023.

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Anastasios N Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction. *arXiv preprint arXiv:2411.11824*, 2024.

Adam Barla, Emanuele Nevali, Luca Viano, and Volkan Cevher. Provably avoiding over-optimization in direct preference optimization without knowing the data distribution. *arXiv preprint arXiv:2602.06239*, 2026.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497 – 1537, 2005.

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.

Tiejin Chen, Xiaoou Liu, Vishnu Nandam, Kuan-Ru Liou, and Hua Wei. Conformal feedback alignment: Quantifying answer-level reliability for robust llm alignment. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 3561–3572, 2026.

Ziyi Chen, Junyi Li, Peiran Yu, and Heng Huang. Provably mitigating corruption, overoptimization, and verbosity simultaneously in offline and online rlhf/dpo alignment. *arXiv preprint arXiv:2510.05526*, 2025.

Prasenjit Dey, Srujana Merugu, and Sivaramakrishnan R Kaveri. Conformal prediction sets for ordinal classification. *Advances in Neural Information Processing Systems*, 36:879–899, 2023.

Tiffany Ding, Anastasios Angelopoulos, Stephen Bates, Michael Jordan, and Ryan J Tibshirani. Class-conditional conformal prediction with many classes. *Advances in neural information processing systems*, 36:64555–64576, 2023.

Jean-Baptiste Fermanian, Pierre Humbert, and Gilles Blanchard. Transductive conformal inference for full ranking. *Advances in Neural Information Processing Systems*, 38:6061–6093, 2025.

Feichen Gan, Youcun Lu, Yingying Zhang, and Yukun Liu. Conformal prediction beyond the horizon: Distribution-free inference for policy evaluation. *Advances in Neural Information Processing Systems*, 38:140902–140917, 2025.

Haosen Ge, Hamsa Bastani, and Osbert Bastani. Stochastic online conformal prediction with semi-bandit feedback. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 18954–18965. PMLR, 2025.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International conference on artificial intelligence and statistics*, volume 238, pp. 4447–4455. PMLR, 2024.

Dhawal Gupta, Adam Fisch, Christoph Dann, and Alekh Agarwal. Mitigating preference hacking in policy optimization with pessimism. *arXiv preprint arXiv:2503.06810*, 2025.

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

Audrey Huang, Wenhao Zhan, Tengyang Xie, Jason Lee, Wen Sun, Akshay Krishnamurthy, and Dylan Foster. Correcting the mythos of kl-regularization: Direct alignment without overoptimization via chi-squared preference optimization. In *International Conference on Learning Representations*, pp. 92647–92701, 2025.

Jianguo Huang, Huajun Xi, Linjun Zhang, Huaxiu Yao, Yue Qiu, and Hongxin Wei. Conformal prediction for deep classifier via label ranking. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 20331–20347. PMLR, 2024.

Xiang Ji, Sanjeev Kulkarni, Mengdi Wang, and Tengyang Xie. Self-play with adversarial critic: Provable and scalable offline alignment for language models. *arXiv preprint arXiv:2406.04274*, 2024.

Cheol Woo Kim, Shresth Verma, Mauricio Tec, and Milind Tambe. Lightweight robust direct preference optimization. *arXiv preprint arXiv:2510.23590*, 2025.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018.

Xize Liang, Chao Chen, Shuang Qiu, Jie Wang, Yue Wu, Zhihang Fu, Hanzhu Chen, Feng Wu, and Jieping Ye. Ropo: Robust preference optimization for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 37131–37161. PMLR, 2025.

Wenbo Liao, Huipeng Huang, Chen Jia, Huajun Xi, Hao Zeng, and Hongxin Wei. Distribution-informed efficient conformal prediction for full ranking. *arXiv preprint arXiv:2601.23128*, 2026.

Zhihan Liu, Miao Lu, Shenao Zhang, Boyi Liu, Hongyi Guo, Yingxiang Yang, Jose Blanchet, and Zhaoran Wang. Provably mitigating overoptimization in rlhf: Your sft loss is implicitly an adversarial regularizer. *Advances in Neural Information Processing Systems*, 37:138663–138697, 2024.

R Duncan Luce. *Individual choice behavior: A theoretical analysis*. Courier Corporation, 2005.

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

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

Licheng Pan, Haocheng Yang, Haoxuan Li, Yichen Sun, Yunsheng Lu, Shijian Wang, Lei Shen, Yuan Lu, Zhixuan Chu, and Hao Wang. Uncertainty-aware reward modeling for stable rlhf. *arXiv preprint arXiv:2606.19818*, 2026.

Yash Patel, Declan Mcnamara, Jackson Loper, Jeffrey Regier, and Ambuj Tewari. Variational inference with coverage guarantees in simulation-based inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 39861–39888, 2024.

Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202, 1975.

Drew Prinster, Clara Fannjiang, Ji Won Park, Kyunghyun Cho, Anqi Liu, Suchi Saria, and Samuel Stanton. Conformal policy control. *arXiv preprint arXiv:2603.02196*, 2026.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sikchi, Joey Hejna, W Bradley Knox, Chelsea Finn, and Scott Niekum. Scaling laws for reward model overoptimization in direct alignment algorithms. *Advances in Neural Information Processing Systems*, 37:126207–126242, 2024.

Sayak Ray Chowdhury, Anush Kini, and Nagarajan Natarajan. Provably robust DPO: Aligning language models with noisy feedback. In *International Conference on Machine Learning*, volume 235, pp. 42258–42274. PMLR, 2024.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of machine learning research*, 9(3):371–421, 2008.

Wenwen Si, Sooyong Jang, Insup Lee, and Osbert Bastani. Conformal constrained policy optimization for cost-effective llm agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 25446–25453, 2026.

Muhammad Faaiz Taufiq, Jean-Francois Ton, Rob Cornish, Yee Whye Teh, and Arnaud Doucet. Conformal off-policy prediction in contextual bandits. *Advances in Neural Information Processing Systems*, 35:31512–31524, 2022.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, second edition, 2022.

Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jiawei Chen, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. Towards robust alignment of language models: Distributionally robustifying direct preference optimization. In *International Conference on Learning Representations*, pp. 16123–16150, 2025a.

Xudong Wu, Jian Qian, Pangpang Liu, Vaneet Aggarwal, and Jiayu Chen. Distributionally robust listwise preference optimization. *arXiv preprint arXiv:2607.01715*, 2026.

Yulian Wu, Rushil Thareja, Praneeth Vepakomma, and Francesco Orabona. Offline and online kl-regularized rlhf under differential privacy. *arXiv preprint arXiv:2510.13512*, 2025b.

Yunpeng Xu, Mufang Ying, Wenge Guo, and Zhi Wei. Two-stage risk control with application to ranked retrieval. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 9104–9111, 2025a.

Zaiyan Xu, Sushil Vemuri, Kishan Panaganti, Dileep Kalathil, Rahul Jain, and Deepak Ramachandran. Robust llm alignment via distributionally robust direct preference optimization. *Advances in Neural Information Processing Systems*, 38:27039–27076, 2025b.

Christopher Yeh, Nicolas Christianson, Alan Wu, Adam Wierman, and Yisong Yue. End-to-end conformal calibration for optimization under uncertainty. *arXiv preprint arXiv:2409.20534*, 2024.

Wenhao Zhan, Masatoshi Uehara, Nathan Kallus, Jason Lee, and Wen Sun. Provable offline preference-based reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 29559–29584, 2024.

Yingying Zhang, Chengchun Shi, and Shikai Luo. Conformal off-policy prediction. In *International Conference on Artificial Intelligence and Statistics*, pp. 2751–2768. PMLR, 2023.

Heyang Zhao, Chenlu Ye, Quanquan Gu, and Tong Zhang. Sharp analysis for kl-regularized contextual bandits and rlhf. *Advances in Neural Information Processing Systems*, 38:107964–108002, 2025.

Minxing Zheng, Wenbin Zhou, and Shixiang Zhu. Beyond prediction: Conformal inference for latent distributional parameters. *arXiv preprint arXiv:2608.03607v2*, 2026.

Xingyu Zhou, Yulian Wu, and Francesco Orabona. A unified theoretical analysis of private and robust offline alignment: from RLHF to DPO. In *International Conference on Machine Learning*, volume 267, pp. 79366–79388. PMLR, 2025.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, and Jiantao Jiao. Starling-7b: Improving llm helpfulness & harmlessness with rlaif, November 2023a.

Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, volume 202, pp. 43037–43067. PMLR, 2023b.

## A Additional Related Work

**Additional Background on Provable and Pessimistic Preference Optimization** Ray Chowdhury et al. (2024) develop a noise-corrected DPO objective with guarantees under random preference-label flips at a known rate, and Zhou et al. (2025) analyze the interplay between privacy and adversarial label corruption in offline RLHF and DPO under linear modeling assumptions. Further pessimistic approaches address general pairwise preference models (Gupta et al., 2025) and combine pessimism with corruption modeling and length regularization in offline alignment (Chen et al., 2025).

**Additional background on conformal prediction.** Conformal prediction constructs distribution-free prediction sets with finite-sample marginal coverage under exchangeability (Vovk et al., 2022; Shafer & Vovk, 2008; Angelopoulos & Bates, 2023; Angelopoulos et al., 2024). Applications include regression (Lei et al., 2018), classification (Ding et al., 2023; Dey et al., 2023; Huang et al., 2024), ranking inference (Fermanian et al., 2025; Liao et al., 2026), and risk control for ranked retrieval (Xu et al., 2025a). Conformal methods have also been developed for off-policy outcome prediction in contextual bandits (Taufiq et al., 2022) and sequential decision-making settings (Zhang et al., 2023), as well as return prediction for infinite-horizon policy evaluation (Gan et al., 2025). Other applications include online prediction with semi-bandit feedback (Ge et al., 2025) and cost-sensitive LLM-agent orchestration under coverage constraints (Si et al., 2026).

## B Validity of Latent Conformal Uncertainty Sets

**Proof of Theorem 1.** Let $R_\gamma = 1 - p_\gamma(\boldsymbol{u}^\star_{n+1}, D_{n+1}) \in [0,1]$. Under the sampling and model-correctness assumptions,

$$\mathbb{E}[R_\gamma] = \mathbb{P}\left\{y_{n+1} \notin \widehat{\mathcal{C}}_\gamma(D_{n+1})\right\} \leq \gamma.$$

The definition of the uncertainty set, together with the empty-set convention, implies

$$\left\{\boldsymbol{u}^\star_{n+1} \notin \widehat{\mathcal{U}}_\alpha(D_{n+1};\gamma)\right\} \subseteq \left\{R_\gamma > \frac{\gamma}{\alpha}\right\}.$$

Markov's inequality therefore yields

$$\mathbb{P}\left\{\boldsymbol{u}^\star_{n+1} \notin \widehat{\mathcal{U}}_\alpha(D_{n+1};\gamma)\right\} \leq \mathbb{P}\left\{R_\gamma > \frac{\gamma}{\alpha}\right\} \leq \frac{\alpha}{\gamma}\mathbb{E}[R_\gamma] \leq \alpha,$$

which completes the proof. □

**Algorithm 1** LatentCP-DPO

**Require:** Labeled preference units $\mathcal{D}_{\mathrm{lab}} = \{(D_i, y_i)\}_{i=1}^{n}$; label-free contexts $\{D_j^{\mathrm{opt}}\}_{j=1}^{n_{\mathrm{opt}}}$; feedback model $\mathrm{M} \in \{\mathrm{BT}, \mathrm{PL}\}$; levels $0 < \gamma < \alpha < 1$.
**Ensure:** LatentCP-DPO policy $\widehat{\pi}_{\mathrm{LCP}}$.
1: Split $\{1, \dots, n\} = \mathcal{I}_{\mathrm{pil}} \dot{\cup} \mathcal{I}_{\mathrm{cal}}$.
2: Fit the pilot policy $\widehat{\pi}_{\mathrm{pil}}$ by solving (5) on the pilot subset.
3: Compute $S_i^{\mathrm{cal}} \leftarrow \widehat{s}(D_i, y_i)$ for all $i \in \mathcal{I}_{\mathrm{cal}}$ using (6), and sort them as $S_{(1)}^{\mathrm{cal}} \le \cdots \le S_{(n_{\mathrm{cal}})}^{\mathrm{cal}}$.
4: Set $k_\gamma \leftarrow \lceil (n_{\mathrm{cal}} + 1)(1 - \gamma) \rceil$.
5: Compute $\widehat{q}_{1-\gamma}^{\mathrm{cal}} \leftarrow S_{(k_\gamma)}^{\mathrm{cal}}$, with $S_{(n_{\mathrm{cal}}+1)}^{\mathrm{cal}} = +\infty$.
6: **for** $i = 1, \dots, n_{\mathrm{opt}}$ **do**
7: Construct $\widehat{\mathcal{C}}_{\gamma,i} \leftarrow \{y \in \mathcal{Y}(D_i^{\mathrm{opt}}) : \widehat{s}(D_i^{\mathrm{opt}}, y) \le \widehat{q}_{1-\gamma}^{\mathrm{cal}}\}$.
8: Construct $\widehat{\mathcal{U}}_{\alpha,i} \leftarrow \{\boldsymbol{u} \in \Theta(D_i^{\mathrm{opt}}) : p_\gamma(\boldsymbol{u}, D_i^{\mathrm{opt}}) \ge 1 - \varepsilon_{\alpha,\gamma}\}$.
9: **if** $\widehat{\mathcal{U}}_{\alpha,i} = \varnothing$ **then**
10: $\widehat{\mathcal{U}}_{\alpha,i} \leftarrow \Theta(D_i^{\mathrm{opt}})$.
11: **end if**
12: For $\pi \in \Pi$, define $\mathcal{L}_{\alpha,\gamma}^{\mathrm{pess}}(\pi; D_i^{\mathrm{opt}})$ according to (11).
13: **end for**
14: Obtain $\widehat{\pi}_{\mathrm{LCP}} \in \operatorname{argmax}_{\pi \in \Pi} \frac{1}{n_{\mathrm{opt}}} \sum_{i=1}^{n_{\mathrm{opt}}} \mathcal{L}_{\alpha,\gamma}^{\mathrm{pess}}(\pi; D_i^{\mathrm{opt}})$.
15: **return** $\widehat{\pi}_{\mathrm{LCP}}$.

# C Additional Details for LatentCP-DPO

## C.1 Overall procedure.

The LatentCP-DPO procedure is summarized in Algorithm 1.

## C.2 BT Instantiation

**Derivation of the Latent Uncertainty Set (12).** The BT likelihood induced by $\boldsymbol{u}$ can be written as

$$p_{\mathrm{BT},\boldsymbol{u}}(y \mid D) = \sigma\left(u^1 - u^0\right)^y \cdot \sigma\left(u^0 - u^1\right)^{1-y}, \qquad y \in \{0, 1\}.$$

For any $\boldsymbol{u} \in \Theta(D)$, let $q_{\boldsymbol{u}} = \sigma\left(u^1 - u^0\right)$. The forward $\gamma$-compatibility (9) can be written as

$$p_\gamma(\boldsymbol{u}, D) = q_{\boldsymbol{u}} \mathbb{I}\{1 \in \widehat{\mathcal{C}}_\gamma(D)\} + (1 - q_{\boldsymbol{u}}) \mathbb{I}\{0 \in \widehat{\mathcal{C}}_\gamma(D)\}, \tag{19}$$

where $\mathbb{I}\{\cdot\}$ denotes the indicator function. Since $\sigma(t) \ge 1 - \frac{\gamma}{\alpha}$ if and only if $t \ge \log \frac{\alpha - \gamma}{\gamma}$, the definition of the latent uncertainty set in (10) yields (12), using the fact that $p_\gamma(\boldsymbol{u}, D) = 1$ when $\widehat{\mathcal{C}}_\gamma(D) = \{0, 1\}$ and the empty-set convention when $\widehat{\mathcal{C}}_\gamma(D) = \varnothing$.

**Derivation of the Pessimistic Objective (13).** With the notation $z_\pi(D) = r_\pi(s, a^1) - r_\pi(s, a^0)$, we write

$$p_{\mathrm{BT},\pi}(y \mid D) = \sigma(z_\pi(D))^y \cdot \sigma(-z_\pi(D))^{1-y}.$$

The expected BT log-likelihood is thus

$$\begin{aligned} G_{\mathrm{BT}}(\boldsymbol{u}, \pi; D) &= \mathbb{E}_{y \sim P_{\mathrm{BT},\boldsymbol{u}}(\cdot \mid D)} \left[\log p_{\mathrm{BT},\pi}(y \mid D)\right] \\ &= q_{\boldsymbol{u}} \log \sigma(z_\pi(D)) + (1 - q_{\boldsymbol{u}}) \log \sigma(-z_\pi(D)) \\ &= \log \sigma(-z_\pi(D)) + q_{\boldsymbol{u}} z_\pi(D), \end{aligned}$$

where the last equality follows from $\log \sigma(z) - \log \sigma(-z) = z$. Thus, $G_{\mathrm{BT}}(\boldsymbol{u}, \pi; D)$ is affine in $q_{\boldsymbol{u}}$.

Combining (10) and (19) with the empty-set convention shows that the closure of the feasible preference-probability set is

$$\mathrm{cl}\{q_{\boldsymbol{u}} : \boldsymbol{u} \in \widehat{\mathcal{U}}_\alpha(D; \gamma)\} = \begin{cases} [1 - \varepsilon_{\alpha,\gamma}, 1], & \widehat{\mathcal{C}}_\gamma(D) = \{1\}, \\ [0, \varepsilon_{\alpha,\gamma}], & \widehat{\mathcal{C}}_\gamma(D) = \{0\}, \\ [0, 1], & \widehat{\mathcal{C}}_\gamma(D) = \{0, 1\} \text{ or } \varnothing. \end{cases}$$

For fixed $\pi$ and $D$, the affine expression above is continuous on $q \in [0,1]$, so taking the closure does not change the infimum. For any $z \in \mathbb{R}$ and $0 \le \underline{q} \le \overline{q} \le 1$,

$$\min_{q \in [\underline{q}, \overline{q}]} \{\log \sigma(-z) + qz\} = \log \sigma(-z) + \underline{q}[z]_+ - \overline{q}[-z]_+.$$

Applying this identity to the three intervals above yields (13).

**Unit-Level Optima of (13).** Fix a context $D$ and regard $z = z_\pi(D)$ as an unrestricted scalar margin. For the singleton conformal prediction sets, it suffices to consider $\widehat{\mathcal{C}}_\gamma(D) = \{1\}$, since the case $\widehat{\mathcal{C}}\gamma(D) = \{0\}$ follows by symmetry under $z \mapsto -z$. Define

$$\mathcal{L}_{\mathrm{BT},1}(z) = \log \sigma\,(z) - \varepsilon_{\alpha,\gamma}[z]_+ = \begin{cases} \log \sigma\,(z) - \varepsilon_{\alpha,\gamma} z, & z > 0, \\ \log \sigma\,(z)\,, & z \le 0. \end{cases}$$

The function $\mathcal{L}_{\mathrm{BT},1}$ is continuous and strictly concave on $\mathbb{R}$, but nondifferentiable at $z = 0$. For $z \neq 0$, its derivative is

$$\mathcal{L}'_{\mathrm{BT},1}(z) = \begin{cases} \sigma(-z), & z < 0, \\ \sigma(-z) - \varepsilon_{\alpha,\gamma}, & z > 0. \end{cases}$$

In particular, its one-sided derivatives at the origin are

$$\mathcal{L}'_{\mathrm{BT},1}(0^-) = \frac{1}{2}, \qquad \mathcal{L}'_{\mathrm{BT},1}(0^+) = \frac{1}{2} - \varepsilon_{\alpha,\gamma}.$$

If $\varepsilon_{\alpha,\gamma} < 1/2$, the unique stationary point on $(0,\infty)$ is the unique global maximizer by strict concavity, with

$$z^\star = \log \frac{1 - \varepsilon_{\alpha,\gamma}}{\varepsilon_{\alpha,\gamma}} > 0, \qquad \sigma(z^\star) = 1 - \varepsilon_{\alpha,\gamma}.$$

If $\varepsilon_{\alpha,\gamma} \ge 1/2$, then $\mathcal{L}_{\mathrm{BT},1}$ is strictly increasing on $(-\infty, 0)$ and decreasing on $(0,\infty)$. Its unique global maximizer is therefore

$$z^\star = 0, \qquad \sigma(z^\star) = \frac{1}{2}.$$

When $\widehat{\mathcal{C}}_\gamma(D) = \{0,1\}$, or when $\widehat{\mathcal{C}}_\gamma(D) = \varnothing$ and the empty-set convention is invoked, the unit-level objective reduces to $\log \sigma(-|z|)$. Since $\log \sigma(t)$ is strictly increasing in $t$,

$$\log \sigma(-|z|) \le \log \sigma(0),$$

with equality if and only if $z = 0$. Hence, the unique global maximizer is again $z^\star = 0$.

**Additional Comparison with Standard DPO.** For a binary preference label $y \in \{0,1\}$, the standard unit-level DPO log-likelihood under the BT preference model is

$$\mathcal{L}^{\mathrm{DPO}}(\pi; D, y) = y \log \sigma\big(z_\pi(D)\big) + (1-y) \log \sigma\big(-z_\pi(D)\big).$$

When viewed in isolation as an unconstrained scalar objective, it drives $\sigma(z_\pi(D))$ toward one for $y = 1$ and toward zero for $y = 0$.

When $\widehat{\mathcal{C}}_\gamma(D)$ is a singleton, the pessimistic objective (13) can be written as

$$\mathcal{L}^{\mathrm{pess}}_{\alpha,\gamma}(\pi; D) = \begin{cases} \mathcal{L}^{\mathrm{DPO}}(\pi; D, 1) - \varepsilon_{\alpha,\gamma}[z_\pi(D)]_+, & \widehat{\mathcal{C}}_\gamma(D) = \{1\}, \\ \mathcal{L}^{\mathrm{DPO}}(\pi; D, 0) - \varepsilon_{\alpha,\gamma}[-z_\pi(D)]_+, & \widehat{\mathcal{C}}_\gamma(D) = \{0\}. \end{cases}$$

For $\widehat{\mathcal{C}}_\gamma(D) = \{1\}$, the one-sided penalty vanishes when $z_\pi(D) < 0$, leaving the standard DPO log-likelihood for $y = 1$ and its margin derivative unchanged. When $z_\pi(D) > 0$, the penalty reduces this derivative by $\varepsilon_{\alpha,\gamma}$. Thus, for $\varepsilon_{\alpha,\gamma} < 1/2$, the pessimistic objective retains the supported preference direction while favoring a finite margin determined by the uncertainty tolerance. The case $\widehat{\mathcal{C}}_\gamma(D) = \{0\}$ is symmetric. The additional penalty follows exactly from worst-case optimization over the compatible preference distributions.

### C.3 PL INSTANTIATION

For any $\boldsymbol{u} \in \Theta(D)$, the induced PL likelihood of a complete ranking $y \in \mathfrak{S}_K$ is

$$p_{\mathrm{PL},\boldsymbol{u}}(y \mid D) = \prod_{k=1}^{K-1} \frac{\exp\left(u^{y_k}\right)}{\sum_{\ell=k}^{K} \exp\left(u^{y_\ell}\right)}.$$

The forward $\gamma$-compatibility under the PL model is

$$p_\gamma(\boldsymbol{u}, D) = \sum_{y \in \widehat{\mathcal{C}}_\gamma(D)} p_{\mathrm{PL},\boldsymbol{u}}(y \mid D).$$

This quantity is the total probability mass that the PL distribution induced by $\boldsymbol{u}$ assigns to the conformal prediction set $\widehat{\mathcal{C}}_\gamma(D)$.

Suppose first that $\widehat{\mathcal{C}}_\gamma(D) \neq \varnothing$. The PL latent uncertainty set is

$$\widehat{\mathcal{U}}_\alpha(D;\gamma) = \left\{\boldsymbol{u} \in \Theta(D) : p_\gamma(\boldsymbol{u}, D) \geq 1 - \varepsilon_{\alpha,\gamma}\right\}. \tag{20}$$

This set is nonempty because a PL distribution can assign probability arbitrarily close to one to any prescribed ranking by taking sufficiently large successive score gaps. For $K > 2$, (20) generally defines an implicit set in the $(K-1)$-dimensional relative-reward space. Using the enumeration $\mathfrak{S}_K = \{y^{(1)}, \ldots, y^{(N_K)}\}$, define $q'_m(\boldsymbol{u}) = p_{\mathrm{PL},\boldsymbol{u}}(y^{(m)} \mid D)$ and recall that $c_m = \mathbb{I}\{y^{(m)} \in \widehat{\mathcal{C}}_\gamma(D)\}$. Let $\boldsymbol{q}'(\boldsymbol{u})$ collect the probabilities $q'_m(\boldsymbol{u})$ in the same enumeration order. For $\boldsymbol{u} \in \Theta(D)$, membership in $\widehat{\mathcal{U}}_\alpha(D;\gamma)$ is equivalent to $\boldsymbol{c}^\top \boldsymbol{q}'(\boldsymbol{u}) \geq 1 - \varepsilon_{\alpha,\gamma}$.

For any $\pi \in \Pi$ and $\boldsymbol{u} \in \Theta(D)$, define

$$G_{\mathrm{PL}}(\boldsymbol{u}, \pi; D) = \mathbb{E}_{y \sim P_{\mathrm{PL},\boldsymbol{u}}(\cdot \mid D)}\left[\log p_{\mathrm{PL},\pi}(y \mid D)\right] = \sum_{y \in \mathfrak{S}_K} p_{\mathrm{PL},\boldsymbol{u}}(y \mid D) \log p_{\mathrm{PL},\pi}(y \mid D).$$

Recalling that $h_m(\pi) = \log p_{\mathrm{PL},\pi}(y^{(m)} \mid D)$, we obtain

$$G_{\mathrm{PL}}(\boldsymbol{u}, \pi; D) = \boldsymbol{q}'(\boldsymbol{u})^\top \boldsymbol{h}(\pi).$$

Consequently, (11) can be rewritten as

$$\mathcal{L}^{\mathrm{pess}}_{\alpha,\gamma}(\pi; D) = \inf_{\boldsymbol{u} \in \Theta(D)} \boldsymbol{q}'(\boldsymbol{u})^\top \boldsymbol{h}(\pi) \quad \text{s.t.} \quad \boldsymbol{c}^\top \boldsymbol{q}'(\boldsymbol{u}) \geq 1 - \varepsilon_{\alpha,\gamma} \tag{21}$$

Every $\boldsymbol{u} \in \Theta(D)$ admits the unique representation $\boldsymbol{u} = B_K \boldsymbol{v}$, where $\boldsymbol{v} = B_K^\top \boldsymbol{u} \in \mathbb{R}^{K-1}$. Since $\boldsymbol{q}'(B_K \boldsymbol{v}) = \boldsymbol{q}(\boldsymbol{v})$, this change of coordinates transforms (21) into (14).

When $\widehat{\mathcal{C}}_\gamma(D) = \varnothing$, we have $p_\gamma(\boldsymbol{u}, D) = 0$ for every $\boldsymbol{u} \in \Theta(D)$. The empty-set convention therefore sets $\widehat{\mathcal{U}}_\alpha(D;\gamma) = \Theta(D)$, and the compatibility constraint is omitted from both formulations.

### C.4 APPROXIMATING THE PESSIMISTIC PL OBJECTIVE

The exact reformulation in (14) makes explicit the computational challenge of the PL instantiation. For fixed context $D$ and policy $\pi \in \Pi$, evaluating the exact pessimistic objective amounts to solving a generally nonconvex constrained optimization problem over the latent-score coordinates $\boldsymbol{v} \in \mathbb{R}^{K-1}$. We therefore develop a feasibility-preserving minorization–maximization (MM) procedure for obtaining an approximate solution to this inner problem.

Recall that $N_K = K!$. Write $h_m = h_m(\pi) = \log p_{\mathrm{PL},\pi}(y^{(m)} \mid D)$. Since all ranking log-likelihoods are finite by the definition of $\Pi$ and $K \geq 2$, we have $h_m < 0$ for every $m = 1, \ldots, N_K$. Define the index set

$$\mathcal{I}_{\mathcal{C}} = \{m : c_m = 1\} = \left\{m : y^{(m)} \in \widehat{\mathcal{C}}_\gamma(D)\right\},$$

and let

$$Q_{\mathcal{C}}(\boldsymbol{v}) = \boldsymbol{c}^\top \boldsymbol{q}(\boldsymbol{v}) = \sum_{m \in \mathcal{I}_{\mathcal{C}}} q_m(\boldsymbol{v}).$$

Thus, $Q_{\mathcal{C}}(\boldsymbol{v}) = p_\gamma(B_K\boldsymbol{v}, D)$ is the forward compatibility expressed in the coordinates $\boldsymbol{v}$. Under the conservative empty-set convention, define the feasible region by

$$\mathcal{V} = \begin{cases} \{\boldsymbol{v} \in \mathbb{R}^{K-1} : Q_{\mathcal{C}}(\boldsymbol{v}) \geq 1 - \varepsilon_{\alpha,\gamma}\}, & \mathcal{I}_{\mathcal{C}} \neq \varnothing, \\ \mathbb{R}^{K-1}, & \mathcal{I}_{\mathcal{C}} = \varnothing. \end{cases}$$

Define

$$F(\boldsymbol{v}) = \sum_{m=1}^{N_K} q_m(\boldsymbol{v}) h_m < 0, \qquad A(\boldsymbol{v}) = \log(-F(\boldsymbol{v})).$$

Since $t \mapsto \log(-t)$ is strictly decreasing on $(-\infty, 0)$,

$$\mathcal{L}^{\text{pess}}_{\alpha,\gamma}(\pi; D) = \inf_{\boldsymbol{v}\in\mathcal{V}} F(\boldsymbol{v}) = -\exp\left(\sup_{\boldsymbol{v}\in\mathcal{V}} A(\boldsymbol{v})\right).$$

Thus, minimizing $F$ over $\mathcal{V}$ is equivalent to maximizing $A$ over the same feasible region. We apply MM to this equivalent maximization problem to obtain a feasible approximate solution.

**Concavity of the PL Log-Likelihood.** Under the parameterization $\boldsymbol{u} = B_K\boldsymbol{v}$, the PL log-likelihood of a ranking $y^{(m)} = (y_1^{(m)}, \ldots, y_K^{(m)})$ is

$$\log q_m(\boldsymbol{v}) = \sum_{k=1}^{K-1} \left[ (B_K\boldsymbol{v})_{y_k^{(m)}} - \log \sum_{\ell=k}^{K} \exp\left((B_K\boldsymbol{v})_{y_\ell^{(m)}}\right) \right].$$

Each summand is a linear function minus a log-sum-exp of linear functions. Therefore, $\log q_m(\boldsymbol{v})$ is smooth and concave in $\boldsymbol{v}$. This property underlies the construction of the MM surrogate objective and constraint below.

**Minorization Step.** At a feasible iterate $\boldsymbol{v}^{(t)} \in \mathcal{V}$, define

$$\eta_m^{(t)} = \frac{q_m(\boldsymbol{v}^{(t)}) h_m}{\sum_{m'=1}^{N_K} q_{m'}(\boldsymbol{v}^{(t)}) h_{m'}}, \qquad m = 1, \ldots, N_K. \tag{22}$$

These weights satisfy $\eta_m^{(t)} > 0$ and $\sum_{m=1}^{N_K} \eta_m^{(t)} = 1$. Jensen's inequality then yields

$$A(\boldsymbol{v}) = \log\left(\sum_{m=1}^{N_K} \eta_m^{(t)} \frac{-h_m q_m(\boldsymbol{v})}{\eta_m^{(t)}}\right) \geq \sum_{m=1}^{N_K} \eta_m^{(t)} \left[\log(-h_m) + \log q_m(\boldsymbol{v}) - \log \eta_m^{(t)}\right].$$

Denote the right-hand side by $M(\boldsymbol{v}, \boldsymbol{v}^{(t)})$. Because each $\log q_m(\boldsymbol{v})$ is concave, $M(\boldsymbol{v}, \boldsymbol{v}^{(t)})$ is concave in $\boldsymbol{v}$. The lower bound holds for all $\boldsymbol{v} \in \mathbb{R}^{K-1}$ and is tight at the current iterate:

$$M(\boldsymbol{v}^{(t)}, \boldsymbol{v}^{(t)}) = A(\boldsymbol{v}^{(t)})$$

Thus, $M(\boldsymbol{v}, \boldsymbol{v}^{(t)})$ is a concave minorizer of $A$.

We next construct a concave minorizer of $\log Q_{\mathcal{C}}(\boldsymbol{v})$. Consider the nontrivial case $0 < |\mathcal{I}_{\mathcal{C}}| < N_K$. At the current feasible iterate $\boldsymbol{v}^{(t)}$, define

$$\omega_m^{(t)} = \frac{q_m(\boldsymbol{v}^{(t)})}{Q_{\mathcal{C}}(\boldsymbol{v}^{(t)})}, \qquad m \in \mathcal{I}_{\mathcal{C}}. \tag{23}$$

These weights satisfy $\omega_m^{(t)} > 0$ for every $m \in \mathcal{I}_{\mathcal{C}}$ and $\sum_{m\in\mathcal{I}_{\mathcal{C}}} \omega_m^{(t)} = 1$. A second application of Jensen's inequality yields

$$\log Q_{\mathcal{C}}(\boldsymbol{v}) = \log\left(\sum_{m\in\mathcal{I}_{\mathcal{C}}} \omega_m^{(t)} \frac{q_m(\boldsymbol{v})}{\omega_m^{(t)}}\right) \geq \sum_{m\in\mathcal{I}_{\mathcal{C}}} \omega_m^{(t)} \left[\log q_m(\boldsymbol{v}) - \log \omega_m^{(t)}\right].$$

Denote the right-hand side by $\psi(\boldsymbol{v}, \boldsymbol{v}^{(t)})$. Since each $\log q_m(\boldsymbol{v})$ is concave, $\psi(\boldsymbol{v}, \boldsymbol{v}^{(t)})$ is concave in $\boldsymbol{v}$, so its superlevel sets are convex. The lower bound holds for all $\boldsymbol{v} \in \mathbb{R}^{K-1}$ and is tight at the current iterate $\boldsymbol{v}^{(t)}$. Consequently,

$$\psi(\boldsymbol{v}, \boldsymbol{v}^{(t)}) \geq \log(1 - \varepsilon_{\alpha,\gamma}) \quad \Rightarrow \quad Q_{\mathcal{C}}(\boldsymbol{v}) \geq 1 - \varepsilon_{\alpha,\gamma}.$$

Moreover, feasibility of $\boldsymbol{v}^{(t)}$ and tightness of the minorizer imply

$$\psi(\boldsymbol{v}^{(t)}, \boldsymbol{v}^{(t)}) = \log Q_{\mathcal{C}}(\boldsymbol{v}^{(t)}) \geq \log(1 - \varepsilon_{\alpha,\gamma}).$$

Thus, the surrogate feasible set is convex, contains the current iterate, and is contained in the original feasible region.

If $\mathcal{I}_{\mathcal{C}} = \varnothing$, the compatibility constraint is omitted under the empty-set convention. If $\mathcal{I}_{\mathcal{C}} = \{1, \ldots, N_K\}$, then $Q_{\mathcal{C}}(\boldsymbol{v}) = 1$ for every $\boldsymbol{v}$, so the constraint is automatically satisfied and may again be omitted.

**Maximization Step.** In the nontrivial constrained case, holding the objective and compatibility weights fixed at iteration $t$, we update the latent-score coordinates by maximizing the concave objective minorizer over the surrogate feasible set. Specifically,

$$\boldsymbol{v}^{(t+1)} \in \operatorname*{argmax}_{\boldsymbol{v} \in \mathbb{R}^{K-1}} M(\boldsymbol{v}, \boldsymbol{v}^{(t)}) \quad \text{s.t.} \quad \psi(\boldsymbol{v}, \boldsymbol{v}^{(t)}) \geq \log(1 - \varepsilon_{\alpha,\gamma}).$$

Dropping the terms in $M(\boldsymbol{v}, \boldsymbol{v}^{(t)})$ that are constant with respect to $\boldsymbol{v}$ and expanding the surrogate constraint gives the equivalent update

$$\begin{aligned} \boldsymbol{v}^{(t+1)} \in \operatorname*{argmax}_{\boldsymbol{v} \in \mathbb{R}^{K-1}} & \sum_{m=1}^{N_K} \eta_m^{(t)} \log q_m(\boldsymbol{v}) \\ \text{s.t.} & \sum_{m \in \mathcal{I}_{\mathcal{C}}} \omega_m^{(t)} \log q_m(\boldsymbol{v}) \geq \log(1 - \varepsilon_{\alpha,\gamma}) + \sum_{m \in \mathcal{I}_{\mathcal{C}}} \omega_m^{(t)} \log \omega_m^{(t)}. \end{aligned} \tag{24}$$

This is a convex optimization problem: the objective is concave in $\boldsymbol{v}$, and the feasible set is convex because it is a superlevel set of a concave function.

**Feasible Initialization.** The MM updates preserve feasibility but require a feasible starting point. The following lemma provides an explicit construction, with the proof provided in Appendix C.5.

**Lemma 1.** *Suppose that $\mathcal{I}_{\mathcal{C}} \neq \varnothing$. Choose any ranking $y^{(m_0)} \in \widehat{\mathcal{C}}_\gamma(D)$ and define $\boldsymbol{u}^{(0)} \in \mathbb{R}^K$ by setting*

$$u^{(0), y_k^{(m_0)}} = \lambda \left( \frac{K+1}{2} - k \right), \qquad k = 1, \ldots, K,$$

*where*

$$\lambda \geq \left[ \log \frac{K-1}{(1 - \varepsilon_{\alpha,\gamma})^{-1/(K-1)} - 1} \right]_+ .$$

*Then $\boldsymbol{u}^{(0)} \in \mathbf{1}_K^\perp$ and $\boldsymbol{v}^{(0)} = B_K^\top \boldsymbol{u}^{(0)}$ satisfies $Q_{\mathcal{C}}(\boldsymbol{v}^{(0)}) \geq 1 - \varepsilon_{\alpha,\gamma}$. Hence, $\boldsymbol{v}^{(0)} \in \mathcal{V}$.*

When $\mathcal{I}_{\mathcal{C}} = \varnothing$, the compatibility constraint is omitted under the conservative empty-set convention, and one may simply initialize $\boldsymbol{v}^{(0)} = \mathbf{0}$.

The preceding steps are summarized in Algorithm 2. Let $\widetilde{\boldsymbol{v}}$ denote the feasible approximate solution returned by the algorithm. The corresponding MM approximation satisfies $\widetilde{\mathcal{L}}_{\alpha,\gamma}^{\mathrm{pess}}(\pi; D) = \boldsymbol{q}(\widetilde{\boldsymbol{v}})^\top \boldsymbol{h}(\pi)$.

**Theoretical Properties.** We next establish the main properties of the proposed MM algorithm, with the proof provided in Appendix C.5.

**Theorem 4** (Feasibility, monotonicity, and stationary accumulation points)**.** *Fix a context $D$ and a policy $\pi$. Suppose that $\boldsymbol{v}^{(0)} \in \mathcal{V}$ and that every MM subproblem is solved exactly. Then*

$$\boldsymbol{v}^{(t)} \in \mathcal{V} \qquad \textit{for every } t \geq 0,$$

*and*

$$A(\boldsymbol{v}^{(t+1)}) \geq A(\boldsymbol{v}^{(t)}), \qquad F(\boldsymbol{v}^{(t+1)}) \leq F(\boldsymbol{v}^{(t)}).$$

*Moreover, both $\{A(\boldsymbol{v}^{(t)})\}_{t \geq 0}$ and $\{F(\boldsymbol{v}^{(t)})\}_{t \geq 0}$ converge to finite limits.*

**Algorithm 2** Feasibility-Preserving MM Algorithm for the Pessimistic PL Inner Problem

**Require:** Fixed context $D$ and policy $\pi$; conformal index set $\mathcal{I}_{\mathcal{C}}$; feasible initialization $\boldsymbol{v}^{(0)} \in \mathcal{V}$; maximum number of iterations $T_{\mathrm{MM}} \geq 1$; tolerance $\delta_{\mathrm{MM}} > 0$.
**Ensure:** A feasible approximate solution $\widetilde{\boldsymbol{v}}$ to the inner problem in (14).

```
Set ṽ ← v^(0).
for t = 0, ..., T_MM − 1 do
    Compute the objective weights {η_m^(t)}_{m=1}^{N_K} using (22).
    if 0 < |I_C| < N_K then
        Compute the compatibility weights {ω_m^(t)}_{m∈I_C} using (23).
        Obtain v^(t+1) by solving the convex subproblem in (24).
    else
        Obtain v^(t+1) by solving the convex subproblem in (24) with the compatibility constraint
        omitted.
    end if
    Set ṽ ← v^(t+1).
    if A(v^(t+1)) − A(v^(t)) ≤ δ_MM then
        break.
    end if
end for
return ṽ.
```

*Suppose, in addition, that $\{\boldsymbol{v}^{(t)}\}_{t\geq 0}$ is bounded, and let $\boldsymbol{v}^\star$ be any of its accumulation points. In the nontrivial constrained case $0 < |\mathcal{I}_{\mathcal{C}}| < N_K$, assume that the Mangasarian–Fromovitz constraint qualification holds at $\boldsymbol{v}^\star$; equivalently,*

$$Q_{\mathcal{C}}(\boldsymbol{v}^\star) = 1 - \varepsilon_{\alpha,\gamma} \quad \Rightarrow \quad \nabla Q_{\mathcal{C}}(\boldsymbol{v}^\star) \neq \mathbf{0}.$$

*Then $\boldsymbol{v}^\star$ satisfies the KKT conditions of the transformed inner maximization problem $\sup_{\boldsymbol{v}\in\mathcal{V}} A(\boldsymbol{v})$ and, equivalently, of the original inner minimization problem in (14).*

The theorem establishes feasibility and monotonic improvement of the inner objective, together with stationarity guarantees under the stated assumptions.

**Computational Complexity.** Consider a fixed context $D$ and policy $\pi$, excluding the cost of evaluating the policy on the $K$ candidates. For a given latent coordinate $\boldsymbol{v}$, computing $B_K\boldsymbol{v}$ costs $O(K^2)$. Given $B_K\boldsymbol{v}$, the PL probability of one ranking can be evaluated in $O(K)$ time using backward cumulative sums, so evaluating all $N_K = K!$ ranking probabilities costs $O(N_K K)$. Once these probabilities are available, updating the weights $\{\eta_m^{(t)}\}_{m=1}^{N_K}$ and $\{\omega_m^{(t)}\}_{m\in\mathcal{I}_{\mathcal{C}}}$ costs $O(N_K)$.

Each MM iteration additionally solves a convex problem in $K-1$ variables with at most one nontrivial scalar inequality constraint. Let $C_{\mathrm{cvx}}(K-1, N_K)$ denote the cost of solving this subproblem to the prescribed tolerance, including evaluations of its surrogate objective and constraint. Up to one-time preprocessing for fixed $(D, \pi)$, an MM run of at most $T_{\mathrm{MM}}$ iterations therefore has time complexity

$$O\left(T_{\mathrm{MM}}\left[N_K K + C_{\mathrm{cvx}}(K-1, N_K)\right]\right).$$

Since $N_K = K!$, explicit enumeration induces factorial dependence on $K$. The proposed implementation is therefore intended for settings with a small-to-moderate number of candidates.

**A Simplex-Relaxation Bound.** A feasible MM output upper-bounds the exact inner infimum. To obtain a complementary lower bound, we relax the PL family of ranking distributions to the full probability simplex while retaining the compatibility constraint. This yields the following computable, generally conservative bound.

**Proposition 1** (Simplex-Relaxation Bound)**.** *Fix a context $D$ and a policy $\pi$, and define*

$$L_{\mathrm{relax}}(\pi; D) = \begin{cases} (1-\varepsilon_{\alpha,\gamma}) \min\limits_{m\in\mathcal{I}_{\mathcal{C}}} h_m(\pi) + \varepsilon_{\alpha,\gamma} \min\limits_{1\leq m\leq N_K} h_m(\pi), & \mathcal{I}_{\mathcal{C}} \neq \varnothing, \\ \min\limits_{1\leq m\leq N_K} h_m(\pi), & \mathcal{I}_{\mathcal{C}} = \varnothing. \end{cases}$$

*For any feasible approximate inner solution $\widetilde{\boldsymbol{v}} \in \mathcal{V}$, let $\widetilde{\mathcal{L}}^{\mathrm{pess}}_{\alpha,\gamma}(\pi; D) = \boldsymbol{q}(\widetilde{\boldsymbol{v}})^\top \boldsymbol{h}(\pi)$. Then*

$$0 \le \widetilde{\mathcal{L}}^{\mathrm{pess}}_{\alpha,\gamma}(\pi; D) - \mathcal{L}^{\mathrm{pess}}_{\alpha,\gamma}(\pi; D) \le \widetilde{\mathcal{L}}^{\mathrm{pess}}_{\alpha,\gamma}(\pi; D) - L_{\mathrm{relax}}(\pi; D). \tag{25}$$

The right-hand side of (25) gives a computable upper bound on the inner value error. This bound can be conservative because the relaxation replaces the $(K-1)$-dimensional PL family by the full $(K!-1)$-dimensional simplex of ranking distributions. The proof is provided in Appendix C.5.

### C.5 Proofs for the PL MM Algorithm

We provide the proofs of Lemma 1, Theorem 4, and Proposition 1 here.

**Proof of Lemma 1.** By construction,

$$\mathbf{1}_K^\top \boldsymbol{u}^{(0)} = \sum_{k=1}^{K} u^{(0),y_k^{(m_0)}} = \lambda \sum_{k=1}^{K} \left( \frac{K+1}{2} - k \right) = 0.$$

Thus, $\boldsymbol{u}^{(0)} \in \mathbf{1}_K^\perp$. Since the columns of $B_K$ form an orthonormal basis of $\mathbf{1}_K^\perp$, setting $\boldsymbol{v}^{(0)} = B_K^\top \boldsymbol{u}^{(0)}$ gives $B_K \boldsymbol{v}^{(0)} = B_K B_K^\top \boldsymbol{u}^{(0)} = \boldsymbol{u}^{(0)}$.

We next lower-bound the PL probability of the selected ranking $y^{(m_0)}$. For every $k = 1, \dots, K-1$, its stage-$k$ selection probability satisfies

$$\frac{\exp\left(u^{(0),y_k^{(m_0)}}\right)}{\sum_{\ell=k}^{K} \exp\left(u^{(0),y_\ell^{(m_0)}}\right)} = \frac{1}{1 + \sum_{\ell=k+1}^{K} \exp\left(\lambda\,(k-\ell)\right)} \ge \frac{1}{1 + (K-1)e^{-\lambda}},$$

because $\lambda \ge 0$ and each of the $K - k \le K - 1$ terms in the sum is at most $e^{-\lambda}$. Multiplying these bounds over the $K-1$ nontrivial stages yields

$$q_{m_0}(\boldsymbol{v}^{(0)}) = \prod_{k=1}^{K-1} \frac{\exp\left(u^{(0),y_k^{(m_0)}}\right)}{\sum_{\ell=k}^{K} \exp\left(u^{(0),y_\ell^{(m_0)}}\right)} \ge \left(1 + (K-1)e^{-\lambda}\right)^{-(K-1)} \ge 1 - \varepsilon_{\alpha,\gamma}.$$

The last inequality follows from the prescribed lower bound on $\lambda$.

Finally, since $y^{(m_0)} \in \widehat{\mathcal{C}}_\gamma(D)$, we have $m_0 \in \mathcal{I}_{\mathcal{C}}$ and hence

$$Q_{\mathcal{C}}(\boldsymbol{v}^{(0)}) = \sum_{m \in \mathcal{I}_{\mathcal{C}}} q_m(\boldsymbol{v}^{(0)}) \ge q_{m_0}(\boldsymbol{v}^{(0)}) \ge 1 - \varepsilon_{\alpha,\gamma}.$$

Therefore, $\boldsymbol{v}^{(0)} \in \mathcal{V}$, completing the proof. □

**Proof of Theorem 4.** First consider the nontrivial constrained case $0 < |\mathcal{I}_{\mathcal{C}}| < N_K$. Whenever $\boldsymbol{v}^{(t)} \in \mathcal{V}$, the tightness property of $\psi$ gives

$$\psi(\boldsymbol{v}^{(t)}, \boldsymbol{v}^{(t)}) = \log Q_{\mathcal{C}}(\boldsymbol{v}^{(t)}) \ge \log(1 - \varepsilon_{\alpha,\gamma}).$$

Thus, the current iterate is feasible for the surrogate subproblem. Moreover, $\psi(\boldsymbol{v}, \boldsymbol{v}^{(t)}) \le \log Q_{\mathcal{C}}(\boldsymbol{v})$, so every surrogate-feasible point belongs to $\mathcal{V}$. Starting from $\boldsymbol{v}^{(0)} \in \mathcal{V}$, feasibility of all iterates follows by induction. When the compatibility constraint is omitted, $\mathcal{V} = \mathbb{R}^{K-1}$ and feasibility is immediate.

By exact maximization and the minorization and tightness properties of $M$,

$$A(\boldsymbol{v}^{(t+1)}) \ge M(\boldsymbol{v}^{(t+1)}, \boldsymbol{v}^{(t)}) \ge M(\boldsymbol{v}^{(t)}, \boldsymbol{v}^{(t)}) = A(\boldsymbol{v}^{(t)}).$$

Since $F(\boldsymbol{v}) = -\exp(A(\boldsymbol{v}))$, this implies $F(\boldsymbol{v}^{(t+1)}) \le F(\boldsymbol{v}^{(t)})$. Furthermore, $\boldsymbol{q}(\boldsymbol{v})$ is a probability vector, so

$$\min_m h_m \le F(\boldsymbol{v}) = \boldsymbol{q}(\boldsymbol{v})^\top \boldsymbol{h} \le \max_m h_m < 0.$$

Consequently, $\{F(\boldsymbol{v}^{(t)})\}_{t\geq 0}$ converges to a finite negative limit. By continuity of $t \mapsto \log(-t)$ on $(-\infty, 0)$, $\{A(\boldsymbol{v}^{(t)})\}_{t\geq 0}$ also converges to a finite limit.

It remains to establish stationarity under the additional assumptions of the theorem. Direct calculation gives

$$\nabla_1 M(\boldsymbol{v}, \boldsymbol{v}) = \nabla A(\boldsymbol{v}),$$

and, in the nontrivial constrained case,

$$\nabla_1 \psi(\boldsymbol{v}, \boldsymbol{v}) = \nabla \log Q_{\mathcal{C}}(\boldsymbol{v}),$$

where $\nabla_1$ denotes differentiation with respect to the first argument. By boundedness, the iterates lie in a compact convex set $\mathcal{K} \subseteq \mathbb{R}^{K-1}$. Let $\boldsymbol{p}^{(1)}(\boldsymbol{v}) \in \mathbb{R}^K$ denote the first-stage PL choice-probability vector, with components

$$p_\ell^{(1)}(\boldsymbol{v}) = \frac{\exp((B_K \boldsymbol{v})_\ell)}{\sum_{\ell'=1}^K \exp((B_K \boldsymbol{v})_{\ell'})}, \qquad \ell = 1, \dots, K.$$

For every ranking, the first-stage contribution to $-\nabla^2 \log q_m(\boldsymbol{v})$ is

$$B_K^\top \left[ \mathrm{Diag}\big(\boldsymbol{p}^{(1)}(\boldsymbol{v})\big) - \boldsymbol{p}^{(1)}(\boldsymbol{v}) \boldsymbol{p}^{(1)}(\boldsymbol{v})^\top \right] B_K \succ 0.$$

Indeed, $\boldsymbol{p}^{(1)}(\boldsymbol{v})$ has full support, and the columns of $B_K$ span $\mathbf{1}_K^\perp$. The remaining stages contribute positive semidefinite matrices. Continuity, compactness, and $\sum_{m=1}^{N_K} \eta_m^{(t)} = 1$ therefore imply that the functions $M(\cdot, \boldsymbol{v}^{(t)})$ are uniformly $\mu_{\mathrm{MM}}$-strongly concave on $\mathcal{K}$ for some $\mu_{\mathrm{MM}} > 0$. Exact maximization over the convex surrogate feasible set then gives

$$A(\boldsymbol{v}^{(t+1)}) - A(\boldsymbol{v}^{(t)}) \geq M(\boldsymbol{v}^{(t+1)}, \boldsymbol{v}^{(t)}) - M(\boldsymbol{v}^{(t)}, \boldsymbol{v}^{(t)}) \geq \frac{\mu_{\mathrm{MM}}}{2} \left\| \boldsymbol{v}^{(t+1)} - \boldsymbol{v}^{(t)} \right\|^2.$$

Convergence of the objective values thus implies $\|\boldsymbol{v}^{(t+1)} - \boldsymbol{v}^{(t)}\| \to 0$.

Now take a subsequence such that $\boldsymbol{v}^{(t_j)} \to \boldsymbol{v}^\star$. Then $\boldsymbol{v}^{(t_j+1)} \to \boldsymbol{v}^\star$ as well. Since $\mathcal{V}$ is closed, $\boldsymbol{v}^\star \in \mathcal{V}$. In the nontrivial constrained case, the KKT conditions of the corresponding surrogate subproblems hold, for all sufficiently large $j$, with some multiplier $\lambda_j \geq 0$:

$$\nabla_1 M\big(\boldsymbol{v}^{(t_j+1)}, \boldsymbol{v}^{(t_j)}\big) + \lambda_j \nabla_1 \psi\big(\boldsymbol{v}^{(t_j+1)}, \boldsymbol{v}^{(t_j)}\big) = \mathbf{0},$$
$$\lambda_j \left[ \psi\big(\boldsymbol{v}^{(t_j+1)}, \boldsymbol{v}^{(t_j)}\big) - \log(1 - \varepsilon_{\alpha,\gamma}) \right] = 0.$$

If the original compatibility constraint is inactive at $\boldsymbol{v}^\star$, continuity implies that the surrogate constraint is also inactive for all sufficiently large $j$. Complementary slackness then gives $\lambda_j = 0$. If the original constraint is active at $\boldsymbol{v}^\star$, continuity, first-order tightness, and the assumed constraint qualification imply

$$\nabla_1 \psi\big(\boldsymbol{v}^{(t_j+1)}, \boldsymbol{v}^{(t_j)}\big) \to \nabla \log Q_{\mathcal{C}}(\boldsymbol{v}^\star) \neq \mathbf{0}.$$

Hence, the norms of the surrogate constraint gradients are bounded away from zero for all sufficiently large $j$. In this active-constraint case, since $\nabla_1 M$ is continuous on the compact set $\mathcal{K} \times \mathcal{K}$, there exists $C < \infty$ such that

$$\left\| \nabla_1 M\big(\boldsymbol{v}^{(t_j+1)}, \boldsymbol{v}^{(t_j)}\big) \right\| \leq C.$$

Moreover, the preceding convergence implies that, for some $c > 0$,

$$\left\| \nabla_1 \psi\big(\boldsymbol{v}^{(t_j+1)}, \boldsymbol{v}^{(t_j)}\big) \right\| \geq c$$

for all sufficiently large $j$. Taking norms in the surrogate stationarity condition therefore gives

$$0 \leq \lambda_j \leq \frac{C}{c}.$$

Together with $\lambda_j = 0$ in the inactive-constraint case, this shows that $\{\lambda_j\}$ is bounded in either case.

Passing to a further subsequence if necessary, let $\lambda_j \to \lambda^\star \geq 0$. Continuity and first-order tightness yield

$$\nabla A(\boldsymbol{v}^\star) + \lambda^\star \nabla \log Q_{\mathcal{C}}(\boldsymbol{v}^\star) = \mathbf{0},$$
$$\log Q_{\mathcal{C}}(\boldsymbol{v}^\star) \geq \log(1 - \varepsilon_{\alpha,\gamma}), \qquad \lambda^\star \left[ \log Q_{\mathcal{C}}(\boldsymbol{v}^\star) - \log(1 - \varepsilon_{\alpha,\gamma}) \right] = 0.$$

These are the KKT conditions for the logarithmic form of the transformed inner problem. Since $Q_{\mathcal{C}} > 0$, replacing the logarithmic constraint by $Q_{\mathcal{C}} \geq 1 - \varepsilon_{\alpha,\gamma}$ only rescales the multiplier by a positive factor.

Finally, $A = \log(-F)$ and $F < 0$ imply $\nabla A = \nabla F / F$. The limiting stationarity condition can therefore be written as

$$\nabla F(\boldsymbol{v}^\star) - \frac{-F(\boldsymbol{v}^\star)\lambda^\star}{Q_{\mathcal{C}}(\boldsymbol{v}^\star)} \nabla Q_{\mathcal{C}}(\boldsymbol{v}^\star) = \mathbf{0}.$$

The rescaled multiplier is nonnegative, and complementary slackness is preserved. Thus, $\boldsymbol{v}^\star$ also satisfies the KKT conditions of the original inner minimization problem.

In the unconstrained cases, exact maximization gives

$$\nabla_1 M\big(\boldsymbol{v}^{(t_j+1)}, \boldsymbol{v}^{(t_j)}\big) = \mathbf{0}.$$

Passing to the limit and using first-order tightness yields $\nabla A(\boldsymbol{v}^\star) = \mathbf{0}$, and hence $\nabla F(\boldsymbol{v}^\star) = \mathbf{0}$. This completes the proof. □

**Proof of Proposition 1.** Fix $D$ and $\pi$, and use the enumeration $\mathfrak{S}_K = \{y^{(1)}, \ldots, y^{(N_K)}\}$ and the notation $q_m(\boldsymbol{v})$, $h_m(\pi)$, $\mathcal{I}_{\mathcal{C}}$, $Q_{\mathcal{C}}(\boldsymbol{v})$, $\mathcal{V}$, and $F(\boldsymbol{v})$ defined above.

If $\mathcal{I}_{\mathcal{C}} \neq \varnothing$, then every $\boldsymbol{v} \in \mathcal{V}$ satisfies

$$\begin{aligned} F(\boldsymbol{v}) &= \sum_{m=1}^{N_K} q_m(\boldsymbol{v}) h_m(\pi) \\ &\geq Q_{\mathcal{C}}(\boldsymbol{v}) \min_{m \in \mathcal{I}_{\mathcal{C}}} h_m(\pi) + (1 - Q_{\mathcal{C}}(\boldsymbol{v})) \min_{1 \leq m \leq N_K} h_m(\pi) \\ &\geq (1 - \varepsilon_{\alpha,\gamma}) \min_{m \in \mathcal{I}_{\mathcal{C}}} h_m(\pi) + \varepsilon_{\alpha,\gamma} \min_{1 \leq m \leq N_K} h_m(\pi) = L_{\mathrm{relax}}(\pi; D). \end{aligned}$$

The second inequality follows from $Q_{\mathcal{C}}(\boldsymbol{v}) \geq 1 - \varepsilon_{\alpha,\gamma}$ and $\min_{m \in \mathcal{I}_{\mathcal{C}}} h_m(\pi) \geq \min_{1 \leq m \leq N_K} h_m(\pi)$. If $\mathcal{I}_{\mathcal{C}} = \varnothing$, the compatibility constraint is omitted, and $F(\boldsymbol{v}) \geq \min_{1 \leq m \leq N_K} h_m(\pi) = L_{\mathrm{relax}}(\pi; D)$. Taking the infimum over $\boldsymbol{v} \in \mathcal{V}$ gives $L_{\mathrm{relax}}(\pi; D) \leq \mathcal{L}^{\mathrm{pess}}_{\alpha,\gamma}(\pi; D)$. Since $\widetilde{\boldsymbol{v}} \in \mathcal{V}$,

$$\mathcal{L}^{\mathrm{pess}}_{\alpha,\gamma}(\pi; D) \leq F(\widetilde{\boldsymbol{v}}) = \widetilde{\mathcal{L}}^{\mathrm{pess}}_{\alpha,\gamma}(\pi; D).$$

Combining these inequalities proves (25). □

# D Suboptimality Upper Bound

## D.1 Full Theorem Statement and Discussion

Recall the conformal uncertainty envelope

$$\Psi_{\alpha,\gamma}(\delta) = \min\left\{1,\, \varepsilon^2_{\alpha,\gamma} + (1 - \varepsilon^2_{\alpha,\gamma})\left(\widetilde{\eta}_{\alpha,\gamma} + \gamma + \frac{\log(2/\delta)}{n_{\mathrm{cal}}}\right)\right\},$$

and the combined error term

$$\mathfrak{E}_{\alpha,\gamma}(\delta) = \underbrace{\mathcal{E}_\Pi}_{\text{policy approximation error}} + \underbrace{\Psi_{\alpha,\gamma}(\delta)}_{\text{conformal uncertainty}} + \underbrace{\frac{d_\Pi + \log(2/\delta)}{n_{\mathrm{opt}}}}_{\text{statistical estimation error}}. \tag{16}$$

We now give the full statement of Theorem 2, including explicit model-specific constants.

**Theorem 2** (Finite-Sample Suboptimality Guarantee, full statement)**.** *Suppose that Assumptions 1 and 2 hold. Assume that the exact empirical LatentCP-DPO objective attains its maximum over* $\Pi$*, and that the localized-complexity condition*

$$\mathfrak{R}_{n_{\mathrm{opt}}}(\zeta) \leq C_{\mathrm{loc}}\left(\sqrt{\frac{d_\Pi \zeta}{n_{\mathrm{opt}}}} + \frac{d_\Pi}{n_{\mathrm{opt}}}\right), \qquad \zeta > 0 \tag{26}$$

*holds for some $C_{\mathrm{loc}} \geq 1$ and $d_\Pi \geq 1$ independent of $n_{\mathrm{opt}}$, where $\mathfrak{R}_{n_{\mathrm{opt}}}(\zeta)$ is defined in the next subsection. Assume further that $C_{\mathrm{comp}}(\pi^\star;\Pi) < \infty$, and that $\widehat{\pi}_{\mathrm{LCP}}$ shares the response support of $\pi^\star$ for $\rho$-almost every prompt. Fix $0 < \gamma < \alpha < 1$ independently of the calibration sample, and let $\delta \in (0,1)$. Conditional on the pilot sample, with probability at least $1-\delta$ over the calibration and optimization samples, every comparator policy $\pi^\dagger$ for which $J(\pi)$ and $H(\pi)$ are finite for $\pi \in \{\pi^\star, \widehat{\pi}_{\mathrm{LCP}}, \pi^\dagger\}$ satisfies*

$$\mathrm{SubOpt}_\beta(\widehat{\pi}_{\mathrm{LCP}}, \pi^\dagger) \leq \frac{C_{\mathrm{M}} C_{\mathrm{comp}}(\pi^\star;\Pi)}{2\beta}\mathfrak{E}_{\alpha,\gamma}(\delta) - \beta\mathbb{E}_{s\sim\rho}\left[D_{\mathrm{KL}}\left(\pi^\dagger(\cdot \mid s)\middle\|\pi^\star(\cdot \mid s)\right)\right]. \quad (17)$$

*Here $C_{\mathrm{M}} > 0$ depends on $B$, the model $\mathrm{M}$, the localized-complexity constant $C_{\mathrm{loc}}$, and, for PL feedback, the fixed number of candidates $K$. For BT feedback, one may take*

$$C_{\mathrm{BT}} = \max\left\{\frac{1}{\sigma(B)\sigma(-B)} + 4,\ \frac{20}{\sigma(B)^2\sigma(-B)^2},\ \frac{4096\log(1+e^B)C_{\mathrm{loc}}^2}{\sigma(B)\sigma(-B)}\right\}, \quad (27)$$

*whereas for PL feedback, one may take*

$$C_{\mathrm{PL}} = \max\Big\{2K(K-1)e^{2B}+4,\ 40K^2(K-1)^2e^{4B},\ 4096Ke^{2B}\big[\log(K!)+BK\big]C_{\mathrm{loc}}^2\Big\}. \quad (28)$$

*In particular, setting $\pi^\dagger = \pi^\star$ removes the comparator-dependent KL correction.*

The error term $\mathfrak{E}_{\alpha,\gamma}(\delta)$ separates three contributions to the upper bound. The policy approximation error $\mathcal{E}_\Pi$ vanishes when $\pi^\star \in \Pi$, because the prompt-dependent normalizing term in (2) cancels from the comparison scores. The conformal uncertainty term $\Psi_{\alpha,\gamma}(\delta)$ accounts for the compatibility tolerance $\varepsilon_{\alpha,\gamma}$, disagreement within the conformal prediction sets, preference-outcome miscoverage, and finite-sample calibration. Finally, $(d_\Pi + \log(2/\delta))/n_{\mathrm{opt}}$ is the estimation contribution from the independent label-free optimization contexts.

The bound also reveals a trade-off in choosing $\gamma$. For fixed pilot scores and calibration data, decreasing $\gamma$ reduces $\varepsilon_{\alpha,\gamma}$ but can only enlarge the conformal prediction sets $\widehat{\mathcal{C}}_\gamma(D)$, potentially leaving more preference comparisons unresolved. Thus, smaller $\gamma$ need not yield a smaller conformal uncertainty term; for the stated guarantee, $\gamma$ must be chosen independently of the calibration data. For fixed $\alpha$, $\gamma$, and $\delta$, the explicit calibration correction in $\Psi_{\alpha,\gamma}(\delta)$ is $O(n_{\mathrm{cal}}^{-1})$, but this is not a rate for the entire conformal term because $\widetilde{\eta}_{\alpha,\gamma}$ also depends on the calibrated sets. Increasing $n_{\mathrm{opt}}$ reduces the optimization-sample contribution, while the approximation and conformal contributions need not vanish.

The explicit constants $C_{\mathrm{BT}}$ and $C_{\mathrm{PL}}$ show how the upper bound depends on the comparison-score range $B$ and, for PL feedback, the number of candidates $K$. A larger $B$ allows more extreme preference probabilities and increases these constants. For fixed $B$ and $C_{\mathrm{loc}}$, the displayed PL constant grows as $O(K^4)$, although the space of complete rankings contains $K!$ outcomes. Polynomial dependence on $K$ also appears in related PL guarantees (Zhu et al., 2023b, Theorem 4.1), although the estimator and performance criterion differ.

In (26), $\mathfrak{R}_{n_{\mathrm{opt}}}(\zeta)$ is the conditional expected Rademacher complexity of the localized excess pessimistic-loss class, defined in (34). Lemma 2 gives sufficient conditions for this bound, covering finite-dimensional affine comparison-score classes (including those induced by log-linear policy classes with a shared finite-dimensional feature map (Zhou et al., 2025)) and uniformly Lipschitz comparison-score parameterizations over bounded parameter domains. Its proof is provided in Appendix D.7.

**Lemma 2** (Sufficient Conditions for Localized Complexity)**.** *Under Assumptions 1 and 2, condition on the pilot and calibration samples, and treat $\Pi$ and its parameterization as fixed under this conditioning. Let $\mathcal{D}_0$ be a set of target contexts with probability one on which the bounds in Assumption 2 hold simultaneously for all $\pi \in \Pi$. Suppose that, for some integer $p \geq 1$, at least one of the following conditions holds:*

1. ***Finite-dimensional comparison scores.*** *The functions $\{D \mapsto x_\pi(D) : \pi \in \Pi\}$, restricted to $\mathcal{D}_0$, lie in an affine function space of dimension at most $p$;*
2. ***Uniformly Lipschitz parameterization.*** *The policy class admits a representation $\Pi = \{\pi_\vartheta : \vartheta \in \mathcal{W}\}$, where $\mathcal{W} \subseteq \mathbb{R}^p$ is bounded and, for some $L_x < \infty$,*

$$\sup_{D\in\mathcal{D}_0}\left\|x_{\pi_\vartheta}(D) - x_{\pi_{\vartheta'}}(D)\right\| \leq L_x\|\vartheta - \vartheta'\|, \qquad \vartheta, \vartheta' \in \mathcal{W}.$$

*Assume that $p$ and the model bounds, together with $\mathrm{diam}(\mathcal{W})$ and $L_x$ in the second case, do not depend on $n_{\mathrm{opt}}$. Then there exists $C_{\mathrm{loc}} \geq 1$, depending only on the model bounds and, in the second case, $L_x \,\mathrm{diam}(\mathcal{W})$, but not on $p$, $n_{\mathrm{opt}}$, or $\zeta$, such that*

$$\mathfrak{R}_{n_{\mathrm{opt}}}(\zeta) \leq C_{\mathrm{loc}} \left( \sqrt{\frac{p\zeta}{n_{\mathrm{opt}}}} + \frac{p}{n_{\mathrm{opt}}} \right), \qquad \zeta > 0.$$

*Consequently, (26) holds with $d_\Pi = p$.*

**Remark 1** (Role of the pilot sample)**.** *Theorem 2 analyzes conformal calibration and pessimistic optimization conditional on the realized pilot sample. Its bound has no separate error term involving $n_{\mathrm{pil}}$. The pilot fit still affects the bound: the nonconformity scores, together with the calibration data, determine the conformal prediction sets, whose informativeness is reflected in $\widetilde{\eta}_{\alpha,\gamma}$. The stated bound requires neither consistency nor a convergence rate for the pilot estimator.*

**Remark 2** (Comparison coefficient for log-linear policies)**.** *The upper bound in Theorem 2 amplifies error by $C_{\mathrm{comp}}(\pi^\star; \Pi)$. This is analogous to the role of relative condition numbers in log-linear offline alignment (Zhou et al., 2025), although the coefficients differ. For BT feedback, suppose that $\Pi = \{\pi_\theta : \theta \in \Theta\}$ has a common log-linear parameterization with bounded feature map $\phi$, that $\pi^\star = \pi_{\theta^\star} \in \Pi$, and that the two comparison responses are drawn conditionally independently from $\pi_{\mathrm{ref}}$. For a fixed evaluation response distribution $\nu$, write $v_\theta = \theta - \theta^\star$. Then*

$$C_{\mathrm{comp}}(\nu; \Pi) = \sup_{\theta : \pi_\theta \in \Pi} \frac{v_\theta^\top \Sigma^{\mathrm{diff}}_{\nu, \pi_\theta} v_\theta}{v_\theta^\top \Sigma^{\mathrm{diff}}_{\pi_{\mathrm{ref}}, \pi_{\mathrm{ref}}} v_\theta},$$

*where*

$$\Sigma^{\mathrm{diff}}_{\nu_1, \nu_2} = \mathbb{E}_{\substack{s \sim \rho \\ a \sim \nu_1(\cdot|s), a' \sim \nu_2(\cdot|s)}} \left[ (\phi(s,a) - \phi(s,a'))(\phi(s,a) - \phi(s,a'))^\top \right],$$

*with $a$ and $a'$ conditionally independent given $s$. The coefficient involves candidate-policy response distributions but evaluates each policy only along its own parameter difference $v_\theta = \theta - \theta^\star$, rather than over arbitrary feature-space directions.*

**Additional Discussion** Theorem 2 concerns an exact empirical optimizer. For a possibly data-dependent output $\widetilde{\pi} \in \Pi$, define the exact empirical pessimistic loss and its optimization gap by

$$\widehat{L}(\pi) = -\frac{1}{n_{\mathrm{opt}}} \sum_{i=1}^{n_{\mathrm{opt}}} \mathcal{L}^{\mathrm{pess}}_{\alpha,\gamma}(\pi; D_i^{\mathrm{opt}}), \qquad \Delta_{\mathrm{opt}} = \widehat{L}(\widetilde{\pi}) - \inf_{\pi \in \Pi} \widehat{L}(\pi) \geq 0.$$

The empirical infimum need not be attained.

**Corollary 1** (Suboptimality under Approximate Empirical Optimization)**.** *Suppose the assumptions of Theorem 2 hold. Conditional on the pilot sample, with probability at least $1 - \delta$ over the calibration and optimization samples, the following bound holds simultaneously for every measurable output $\widetilde{\pi} \in \Pi$ that shares the response support of $\pi^\star$ for $\rho$-almost every prompt and every comparator $\pi^\dagger$ such that $J(\pi)$ and $H(\pi)$ are finite for each $\pi \in \{\pi^\star, \widetilde{\pi}, \pi^\dagger\}$:*

$$\begin{aligned} \mathrm{SubOpt}_\beta(\widetilde{\pi}, \pi^\dagger) \leq & \frac{C_{\mathrm{M}} C_{\mathrm{comp}}(\pi^\star; \Pi)}{2\beta} \mathfrak{E}_{\alpha,\gamma}(\delta) + \frac{3 C_{\mathrm{comp}}(\pi^\star; \Pi)}{\beta \kappa_{\mathrm{M}}} \Delta_{\mathrm{opt}} \\ & - \beta \mathbb{E}_{s \sim \rho} \left[ D_{\mathrm{KL}} \left( \pi^\dagger(\cdot \mid s) \middle\| \pi^\star(\cdot \mid s) \right) \right], \end{aligned} \tag{29}$$

*where $\kappa_{\mathrm{M}}$ is the curvature constant in Lemma 4.*

The additional term measures the gap in the exact empirical objective. The proof is given in Appendix D.6.

For PL feedback, the effect of the MM approximation on $\Delta_{\mathrm{opt}}$ can be made explicit. Let $\widetilde{\boldsymbol{v}}_i$ be a feasible MM output evaluated at $\widetilde{\pi}$ on $D_i^{\mathrm{opt}}$. For each $D_i^{\mathrm{opt}}$, let $\boldsymbol{q}_i$ and $\boldsymbol{h}_i$ denote the vectors in (14) formed using that context's ranking enumeration. Holding the MM outputs fixed as $\pi$ varies, define

$$\widehat{L}_{\mathrm{fr}}(\pi) = -\frac{1}{n_{\mathrm{opt}}} \sum_{i=1}^{n_{\mathrm{opt}}} \boldsymbol{q}_i(\widetilde{\boldsymbol{v}}_i)^\top \boldsymbol{h}_i(\pi),$$

and set

$$\epsilon_{\mathrm{in}} = \widehat{L}(\widetilde{\pi}) - \widehat{L}_{\mathrm{fr}}(\widetilde{\pi}), \qquad \epsilon_{\mathrm{out}} = \widehat{L}_{\mathrm{fr}}(\widetilde{\pi}) - \inf_{\pi\in\Pi} \widehat{L}_{\mathrm{fr}}(\pi).$$

We have $\widehat{L}_{\mathrm{fr}}(\pi) \le \widehat{L}(\pi)$ for every $\pi \in \Pi$. Consequently, Proposition 1 gives

$$\Delta_{\mathrm{opt}} \le \epsilon_{\mathrm{in}} + \epsilon_{\mathrm{out}} \le \epsilon_{\mathrm{out}} + \frac{1}{n_{\mathrm{opt}}} \sum_{i=1}^{n_{\mathrm{opt}}} \left[ \widetilde{\mathcal{L}}^{\mathrm{pess}}_{\alpha,\gamma}(\widetilde{\pi}; D_i^{\mathrm{opt}}) - L_{\mathrm{relax}}(\widetilde{\pi}; D_i^{\mathrm{opt}}) \right]. \tag{30}$$

Thus, $\Delta_{\mathrm{opt}}$ is bounded by the sum of the outer optimization gap and a computable upper bound on the inner approximation error.

**Proof Roadmap** The proof of Theorem 2 proceeds in three ingredients. Appendix D.3 controls the excess pessimistic preference risk of the exact empirical optimizer. Appendix D.4 then uses $C_{\mathrm{comp}}$ to obtain a KL-regularized suboptimality bound in terms of pessimistic benchmark bias, approximation error, and estimation error. Appendix D.5 bounds the bias and approximation terms using conformal uncertainty and finite-sample calibration. Appendix D.6 then combines these results to prove the Theorem 2 and Corollary 1. Proofs of the auxiliary results are collected in Appendix D.7.

### D.2 Setup and Auxiliary Quantities

For the optimization-sample analysis, we condition on the realized pilot and calibration samples. The sets $\widehat{\mathcal{C}}_\gamma(D)$ and $\widehat{\mathcal{U}}_\alpha(D;\gamma)$ are then fixed functions of $D$, while Assumption 1 ensures that $D_1^{\mathrm{opt}}, \dots, D_{n_{\mathrm{opt}}}^{\mathrm{opt}}$ remain i.i.d. draws from the target context distribution. When deriving probability bounds over the calibration sample, we instead condition only on the pilot sample.

For a BT scalar score or a centered PL score $x$, let $p_{\mathrm{M},x}(y \mid D)$ denote the corresponding preference likelihood, and define

$$\ell_x(y, D) = -\log p_{\mathrm{M},x}(y \mid D).$$

The unit-level pessimistic loss is

$$f_D(x) = \sup_{\boldsymbol{u}\in\widehat{\mathcal{U}}_\alpha(D;\gamma)} \mathbb{E}_{y\sim P_{\mathrm{M},\boldsymbol{u}}(\cdot|D)}[\ell_x(y, D)]. \tag{31}$$

By the definition of $x_\pi(D)$, we have $p_{\mathrm{M},x_\pi(D)}(y \mid D) = p_{\mathrm{M},\pi}(y \mid D)$. Thus, $f_D(x_\pi(D)) = -\mathcal{L}^{\mathrm{pess}}_{\alpha,\gamma}(\pi; D)$.

Define the conditional population risk and its empirical counterpart by

$$L(\pi) = \mathbb{E}_D[f_D(x_\pi(D))], \qquad \widehat{L}(\pi) = \frac{1}{n_{\mathrm{opt}}} \sum_{i=1}^{n_{\mathrm{opt}}} f_{D_i^{\mathrm{opt}}}(x_\pi(D_i^{\mathrm{opt}})),$$

where $\mathbb{E}_D$ is over a fresh target context, conditional on the pilot and calibration samples. Under the empirical-attainment condition of Theorem 2, the exact LatentCP-DPO estimator satisfies

$$\widehat{\pi}_{\mathrm{LCP}} \in \operatorname*{argmin}_{\pi\in\Pi} \widehat{L}(\pi). \tag{32}$$

The guarantees below concern exact inner supremum values in (31) and a global empirical minimizer in (32).

### D.3 Excess Preference-Risk Bounds

Define the bounded comparison-score domain by

$$\mathcal{X}(D) = \begin{cases} [-B, B], & \mathrm{M} = \mathrm{BT}, \\ \left\{ x \in \mathbb{R}^K : \mathbf{1}_K^\top x = 0, \|x\|_\infty \le B \right\}, & \mathrm{M} = \mathrm{PL}. \end{cases}$$

By Assumption 2, $\mathcal{X}(D)$ contains $x^\star(D)$ and $x_\pi(D)$ for every $\pi \in \Pi$, almost surely. We first establish uniform bounds on the negative log-likelihood and the corresponding pessimistic loss.

**Lemma 3** (Uniform Boundedness of the Pessimistic Loss)**.** *Under Assumption 2, define*

$$H_{\mathrm{M}} = \begin{cases} \log(1+e^{B}), & \mathrm{M} = \mathrm{BT}, \\ \log(K!) + BK, & \mathrm{M} = \mathrm{PL}. \end{cases}$$

*For almost every target context $D$, every $x \in \mathcal{X}(D)$, and every $y \in \mathcal{Y}(D)$,*

$$0 \le \ell_x(y, D) \le H_{\mathrm{M}}, \qquad 0 \le f_D(x) \le H_{\mathrm{M}}.$$

Having established uniform boundedness, we next relate excess pessimistic loss to the squared distance from a pointwise minimizer.

**Lemma 4** (Pointwise Benchmark and Quadratic Growth)**.** *Under Assumption 2, define*

$$\kappa_{\mathrm{M}} = \begin{cases} \sigma(B)\sigma(-B), & \mathrm{M} = \mathrm{BT}, \\ e^{-2B}/K, & \mathrm{M} = \mathrm{PL}. \end{cases}$$

*For almost every target context $D$, the function $f_D$ is continuous and $\kappa_{\mathrm{M}}$-strongly convex on $\mathcal{X}(D)$. Consequently, it admits a unique minimizer*[4]

$$\bar{x}(D) = \operatorname*{argmin}_{x \in \mathcal{X}(D)} f_D(x).$$

*Moreover, for every $x \in \mathcal{X}(D)$,*

$$\frac{\kappa_{\mathrm{M}}}{2} \|x - \bar{x}(D)\|^2 \le f_D(x) - f_D(\bar{x}(D)) \le H_{\mathrm{M}}. \tag{33}$$

We next control sampling fluctuations uniformly over the policy class. Define the pointwise excess pessimistic loss and its best-in-class population excess pessimistic risk by

$$e_\pi(D) = f_D(x_\pi(D)) - f_D(\bar{x}(D)), \qquad \mathcal{E}_\Pi^{\mathrm{pess}} = \inf_{\pi \in \Pi} \mathbb{E}_D[e_\pi(D)].$$

Lemma 4 implies that $0 \le e_\pi(D) \le H_{\mathrm{M}}$ and hence

$$\mathrm{Var}_D(e_\pi(D)) \le \mathbb{E}_D[e_\pi(D)^2] \le H_{\mathrm{M}} \mathbb{E}_D[e_\pi(D)].$$

To exploit the mean-dependent variance bound, for $\zeta > 0$, define the localized class of rescaled excess-loss functions by

$$\mathcal{H}_\zeta = \left\{ \frac{\lambda e_\pi}{H_{\mathrm{M}}} : \pi \in \Pi, 0 \le \lambda \le 1, \frac{\lambda \mathbb{E}_D[e_\pi(D)]}{H_{\mathrm{M}}} \le \zeta \right\}.$$

Every $h \in \mathcal{H}_\zeta$ takes values in $[0,1]$ and satisfies $\mathrm{Var}_D(h(D)) \le \zeta$. This rescaling is used only for analysis and does not alter the policy class or the training objective.

Define the conditional expected Rademacher complexity of $\mathcal{H}_\zeta$ by

$$\mathfrak{R}_{n_{\mathrm{opt}}}(\zeta) = \mathbb{E}\left[ \sup_{h \in \mathcal{H}_\zeta} \left| \frac{1}{n_{\mathrm{opt}}} \sum_{i=1}^{n_{\mathrm{opt}}} \epsilon_i h(D_i^{\mathrm{opt}}) \right| \right]. \tag{34}$$

Here, $\epsilon_1, \ldots, \epsilon_{n_{\mathrm{opt}}}$ are independent Rademacher signs, each taking the values $+1$ and $-1$ with equal probability, and are independent of the optimization contexts. The expectation is over both the optimization contexts and these signs, conditional on the pilot and calibration samples. We assume that the displayed suprema define random variables, so that their expectations are well defined.

**Lemma 5** (Relative Uniform-Convergence Bound)**.** *Under Assumptions 1 and 2, fix $\delta \in (0,1)$. Let $\zeta_{n_{\mathrm{opt}}} > 0$ be fixed conditional on the pilot and calibration samples and satisfy*

$$16\mathfrak{R}_{n_{\mathrm{opt}}}(\zeta_{n_{\mathrm{opt}}}) \le \zeta_{n_{\mathrm{opt}}}, \qquad \frac{64 \log(1/\delta)}{n_{\mathrm{opt}}} \le \zeta_{n_{\mathrm{opt}}}. \tag{35}$$

[4]The benchmark minimizes the pessimistic loss over the bounded score domain, not over the policy class. It need not equal $x^\star(D)$ or be realizable by any policy in $\Pi$.

*Set $\xi_{n_{\mathrm{opt}}} = H_{\mathrm{M}}\zeta_{n_{\mathrm{opt}}}$. Then, with conditional probability at least $1-\delta$, simultaneously for all $\pi \in \Pi$,*

$$\left| \frac{1}{n_{\mathrm{opt}}} \sum_{i=1}^{n_{\mathrm{opt}}} e_\pi(D_i^{\mathrm{opt}}) - \mathbb{E}_D[e_\pi(D)] \right| \le \frac{1}{3} \max\left\{ \mathbb{E}_D[e_\pi(D)], \xi_{n_{\mathrm{opt}}} \right\}. \tag{36}$$

*On the same event, the exact estimator satisfies*

$$\mathbb{E}_D\left[e_{\widehat{\pi}_{\mathrm{LCP}}}(D)\right] \le \min\left\{ H_{\mathrm{M}}, \mathcal{E}_\Pi^{\mathrm{pess}} + \max\{\mathcal{E}_\Pi^{\mathrm{pess}}, \xi_{n_{\mathrm{opt}}}\} \right\}. \tag{37}$$

*Consequently,*

$$\mathbb{E}_D\left[ \|x_{\widehat{\pi}_{\mathrm{LCP}}}(D) - \bar{x}(D)\|^2 \right] \le \frac{2}{\kappa_{\mathrm{M}}} \min\left\{ H_{\mathrm{M}}, \mathcal{E}_\Pi^{\mathrm{pess}} + \max\{\mathcal{E}_\Pi^{\mathrm{pess}}, \xi_{n_{\mathrm{opt}}}\} \right\}. \tag{38}$$

**Remark 3.** *Since every $h \in \mathcal{H}_\zeta$ takes values in $[0,1]$, we have $\mathfrak{R}_{n_{\mathrm{opt}}}(\zeta) \le 1$. Thus, the conditions in (35) always admit a solution; for example, one may take*

$$\zeta_{n_{\mathrm{opt}}} = \max\left\{ 16, \frac{64\log(1/\delta)}{n_{\mathrm{opt}}} \right\}.$$

*This choice establishes existence but yields only a trivial bound. Under the localized-complexity condition, Corollary 2 gives a radius that tends to zero as $n_{\mathrm{opt}} \to \infty$ for fixed $C_{\mathrm{loc}}$, $d_\Pi$, and $\delta$; Lemma 2 gives sufficient conditions for that complexity bound.*

We now relate the preceding estimation bound to the true preference distribution. Let $P_{\mathrm{M},x}(\cdot \mid D)$ denote the distribution with probability mass function $p_{\mathrm{M},x}(\cdot \mid D)$. For any $\pi \in \Pi$, define the excess preference risk by

$$\mathcal{R}(\pi) = \mathbb{E}_D\left[ D_{\mathrm{KL}}\left( P_{\mathrm{M},x^\star(D)}(\cdot \mid D) \middle\| P_{\mathrm{M},x_\pi(D)}(\cdot \mid D) \right) \right].$$

Under model correctness, this is the excess population negative log-likelihood relative to the true comparison score $x^\star(D)$, with both losses evaluated under the true preference distribution. Since the pessimistic benchmark $\bar{x}(D)$ need not coincide with $x^\star(D)$, define its preference-risk bias by

$$b_{\alpha,\gamma} = \mathbb{E}_D\left[ D_{\mathrm{KL}}\left( P_{\mathrm{M},x^\star(D)}(\cdot \mid D) \middle\| P_{\mathrm{M},\bar{x}(D)}(\cdot \mid D) \right) \right].$$

This bias is fixed conditional on the realized pilot and calibration samples and is not assumed to vanish.

**Theorem 5** (Excess Preference-Risk Bound)**.** *Under the conditions of Lemma 5, define*

$$\Lambda_{\mathrm{M}} = \begin{cases} \dfrac{1}{4}, & \mathrm{M} = \mathrm{BT}, \\ \dfrac{K-1}{2}, & \mathrm{M} = \mathrm{PL}. \end{cases}$$

*Then, with conditional probability at least $1-\delta$,*

$$\sqrt{\mathcal{R}(\widehat{\pi}_{\mathrm{LCP}})} \le \sqrt{b_{\alpha,\gamma}} + \sqrt{\frac{\Lambda_{\mathrm{M}}}{\kappa_{\mathrm{M}}} \min\left\{ H_{\mathrm{M}}, \mathcal{E}_\Pi^{\mathrm{pess}} + \max\left\{ \mathcal{E}_\Pi^{\mathrm{pess}}, \xi_{n_{\mathrm{opt}}} \right\} \right\}}. \tag{39}$$

*Moreover, deterministically, $\mathcal{R}(\pi) \le H_{\mathrm{M}}$ for every $\pi \in \Pi$.*

Theorem 5 bounds the learned policy's excess preference risk in terms of the pessimistic benchmark bias, approximation to the pointwise pessimistic benchmark, and optimization-sample error. The following corollary makes the last term explicit under a localized-complexity condition.

**Corollary 2** (Explicit Statistical Rate under Localized Complexity)**.** *Under the conditions of Theorem 5, suppose that, conditional on the pilot and calibration samples, there exist constants $C_{\mathrm{loc}} \ge 1$ and $d_\Pi \ge 1$, independent of $n_{\mathrm{opt}}$, such that*

$$\mathfrak{R}_{n_{\mathrm{opt}}}(\zeta) \le C_{\mathrm{loc}} \left( \sqrt{\frac{d_\Pi \zeta}{n_{\mathrm{opt}}}} + \frac{d_\Pi}{n_{\mathrm{opt}}} \right), \qquad \zeta > 0 \tag{26}$$

*For any $\delta \in (0,1)$, the choice*

$$\zeta_{n_{\mathrm{opt}}} = \frac{\max\left\{ 1024 C_{\mathrm{loc}}^2 d_\Pi, 64\log(1/\delta) \right\}}{n_{\mathrm{opt}}} \tag{40}$$

*satisfies (35). Consequently, with conditional probability at least* $1-\delta$*,*

$$\mathcal{R}(\widehat{\pi}_{\mathrm{LCP}}) \leq \underbrace{2b_{\alpha,\gamma} + \frac{4\Lambda_{\mathrm{M}}}{\kappa_{\mathrm{M}}}\mathcal{E}_{\Pi}^{\mathrm{pess}}}_{\textit{benchmark bias and approximation}} + \underbrace{\frac{2\Lambda_{\mathrm{M}} H_{\mathrm{M}}}{\kappa_{\mathrm{M}} n_{\mathrm{opt}}} \max\left\{1024 C_{\mathrm{loc}}^2 d_{\Pi}, 64\log(1/\delta)\right\}}_{\textit{statistical error}}. \tag{41}$$

*The right-hand side may additionally be capped at* $H_{\mathrm{M}}$*.*

Conditional on the pilot and calibration samples, condition (26) constrains $\Pi$ through the localized excess pessimistic-loss class it induces. Lemma 2 gives sufficient conditions for this bound.

### D.4 Preference-to-Policy Transfer

We next use $C_{\mathrm{comp}}(\pi^\star;\Pi)$ to transfer comparison-score error to policy KL divergence.

**Lemma 6** (Comparison-Score Error to Policy KL Divergence)**.** *Suppose that Assumption 2 holds and that* $C_{\mathrm{comp}}(\pi^\star;\Pi) < \infty$*. Then, for every* $\pi \in \Pi$ *sharing the response support of* $\pi^\star$ *for* $\rho$*-almost every prompt,*

$$\mathbb{E}_{s\sim\rho}\left[D_{\mathrm{KL}}\left(\pi(\cdot \mid s)\|\pi^\star(\cdot \mid s)\right)\right] \leq \frac{C_{\mathrm{comp}}(\pi^\star;\Pi)}{2\beta^2}\mathbb{E}_D\left[\|x_\pi(D) - x^\star(D)\|^2\right]. \tag{42}$$

We next combine the comparison-score estimation bound with Lemma 6 to control KL-regularized policy suboptimality.

**Theorem 6** (Regularized Suboptimality Bound)**.** *Under the conditions of Lemma 5, let* $\xi_{n_{\mathrm{opt}}}$ *be as defined there. Suppose that the exact empirical LatentCP-DPO objective attains its maximum over* $\Pi$*, and let* $\widehat{\pi}_{\mathrm{LCP}}$ *be any maximizer. Assume that* $C_{\mathrm{comp}}(\pi^\star;\Pi) < \infty$*. Let* $\pi^\dagger$ *be a comparator such that* $J(\pi)$ *and* $H(\pi)$ *are finite for* $\pi \in \{\pi^\star, \widehat{\pi}_{\mathrm{LCP}}, \pi^\dagger\}$*. Conditional on the pilot and calibration samples, there is an event of probability at least* $1-\delta$ *over the optimization sample on which, provided that* $\widehat{\pi}_{\mathrm{LCP}}$ *shares the response support of* $\pi^\star$ *for* $\rho$*-almost every prompt, the following bound holds:*

$$\begin{aligned}\mathrm{SubOpt}_\beta(\widehat{\pi}_{\mathrm{LCP}}, \pi^\dagger) \leq & \frac{C_{\mathrm{comp}}(\pi^\star;\Pi)}{\beta\kappa_{\mathrm{M}}} \min\left\{H_{\mathrm{M}}, \left(\sqrt{b_{\alpha,\gamma}} + \sqrt{\mathcal{E}_{\Pi}^{\mathrm{pess}} + \max\left\{\mathcal{E}_{\Pi}^{\mathrm{pess}}, \xi_{n_{\mathrm{opt}}}\right\}}\right)^2\right\} \\ & - \beta\mathbb{E}_{s\sim\rho}\left[D_{\mathrm{KL}}\left(\pi^\dagger(\cdot \mid s)\middle\|\pi^\star(\cdot \mid s)\right)\right].\end{aligned} \tag{43}$$

To quantify the effect of conformal uncertainty on this bound, we next control $b_{\alpha,\gamma}$ and $\mathcal{E}_{\Pi}^{\mathrm{pess}}$ in terms of the calibrated prediction sets and the policy-class approximation error $\mathcal{E}_{\Pi}$.

### D.5 Effect of Conformal Uncertainty

To quantify the effect of conformal prediction, we first analyze calibration-sample variability conditional only on the pilot sample. For each realized calibration sample, define the conditional preference-outcome miscoverage by

$$\bar{\gamma}_{\mathrm{cal}} = \mathbb{E}_D\left[1 - p_\gamma\big(\boldsymbol{u}^\star(D), D\big)\right],$$

where the expectation is over the target context distribution, with the pilot and calibration samples held fixed. Under model correctness, this quantity equals the probability that a target preference outcome falls outside $\widehat{\mathcal{C}}_\gamma(D)$, conditional on these samples. The following lemma controls this miscoverage in terms of $\gamma$ and the calibration sample size.

**Lemma 7** (Conditional Calibration Miscoverage)**.** *Under Assumption 1, fix* $\gamma \in (0,\alpha)$ *independently of the calibration sample, and let* $n_{\mathrm{cal}} \geq 1$*. Set* $j_\gamma = \lfloor (n_{\mathrm{cal}}+1)\gamma \rfloor$*. For* $\delta_{\mathrm{cal}} \in (0,1)$*, let* $u_{n_{\mathrm{cal}},\gamma}(\delta_{\mathrm{cal}}) = 0$ *if* $j_\gamma = 0$*; otherwise, let it be the* $(1-\delta_{\mathrm{cal}})$*-quantile of* $\mathrm{Beta}(j_\gamma, n_{\mathrm{cal}}+1-j_\gamma)$*. Then, conditional on the pilot sample, with probability at least* $1-\delta_{\mathrm{cal}}$ *over the calibration sample,*

$$\bar{\gamma}_{\mathrm{cal}} \leq u_{n_{\mathrm{cal}},\gamma}(\delta_{\mathrm{cal}}) \leq \min\left\{1, 2\gamma + \frac{4\log(1/\delta_{\mathrm{cal}})}{3n_{\mathrm{cal}}}\right\}. \tag{44}$$

*These bounds allow ties and infinite scores. If the score distribution conditional on the pilot sample is continuous and* $j_\gamma \geq 1$*, then* $\bar{\gamma}_{\mathrm{cal}}$ *has the* $\mathrm{Beta}(j_\gamma, n_{\mathrm{cal}}+1-j_\gamma)$ *distribution conditional on the pilot sample.*

We next relate the miscoverage and informativeness of the conformal prediction sets to the benchmark bias and the policy-class approximation error. Throughout the following analysis, we condition on the realized pilot and calibration samples. Recall

$$\eta_\gamma = \mathbb{P}_D\left\{|\widehat{\mathcal{C}}_\gamma(D)| \neq 1\right\}, \qquad \mathcal{E}_\Pi = \inf_{\pi\in\Pi} \mathbb{E}_D\left[\|x_\pi(D) - x^\star(D)\|^2\right].$$

Here, $\eta_\gamma$ is the probability of a non-singleton or empty conformal prediction set. The quantity $\mathcal{E}_\Pi$ measures the policy class's mean squared approximation error relative to the true comparison score, whereas $\mathcal{E}_\Pi^{\mathrm{pess}}$ measures excess pessimistic risk relative to the benchmark $\bar{x}(D)$.

To retain partial ranking information, for PL feedback, write $R_k(y) = \{y_k, \dots, y_K\}$. Define the normalized set disagreement by

$$\omega_D = \begin{cases} \mathbb{I}\{|\widehat{\mathcal{C}}_\gamma(D)| \neq 1\}, & \mathrm{M} = \mathrm{BT}, \\ 1, & \mathrm{M} = \mathrm{PL},\ \widehat{\mathcal{C}}_\gamma(D) = \varnothing, \\ \frac{1}{K-1}\max_{y,z\in\widehat{\mathcal{C}}_\gamma(D)} \sum_{k=2}^K \mathbb{I}\{R_k(y) \neq R_k(z)\}, & \mathrm{M} = \mathrm{PL},\ \widehat{\mathcal{C}}_\gamma(D) \neq \varnothing. \end{cases}$$

Set

$$r_\varepsilon(w) = \frac{2\varepsilon_{\alpha,\gamma} w + (1-\varepsilon_{\alpha,\gamma})w^2}{1+\varepsilon_{\alpha,\gamma}}, \qquad \widetilde{\eta}_{\alpha,\gamma} = \mathbb{E}_D[r_\varepsilon(\omega_D)]. \tag{45}$$

Since $w^2 \le r_\varepsilon(w) \le w$ for $w \in [0,1]$ and $\omega_D \le \mathbb{I}\{|\widehat{\mathcal{C}}_\gamma(D)| \neq 1\}$, we have $0 \le \widetilde{\eta}_{\alpha,\gamma} \le \eta_\gamma$, and $\widetilde{\eta}_{\alpha,\gamma} = \eta_\gamma$ for BT feedback. For example, the PL set $\widehat{\mathcal{C}}_\gamma(D) = \{(1,2,3),(2,1,3)\}$ has $\omega_D = 1/2$ and contributes $(1+3\varepsilon_{\alpha,\gamma})/(4(1+\varepsilon_{\alpha,\gamma}))$ rather than one. This measure quantifies unresolved ranking structure; a common pairwise order alone does not identify the full preference distribution.

Let $t_D = 1 - p_\gamma(\boldsymbol{u}^\star(D), D)$ and define

$$\Xi_{\alpha,\gamma} = \mathbb{E}_D\left[\left(\omega_D + (1-\omega_D)\max\{\varepsilon_{\alpha,\gamma}, t_D\}\right)^2\right].$$

For singleton sets, the integrand is $\max\{\varepsilon_{\alpha,\gamma}, t_D\}^2$, a squared total-variation envelope; for empty sets, it equals one. For non-singleton PL sets, it instead uses a bound on the difference between expected loss gradients, retaining information shared by the rankings. The following lemma connects this discrepancy envelope to the benchmark bias and pessimistic approximation error.

**Lemma 8** (Conformal Control of Benchmark and Approximation Bias)**.** *Under Assumption 2, let $H_{\mathrm{M}}$, $\kappa_{\mathrm{M}}$, and $\Lambda_{\mathrm{M}}$ be defined as above, and set*

$$G_{\mathrm{M}} = \begin{cases} 1, & \mathrm{M} = \mathrm{BT}, \\ \sqrt{2}(K-1), & \mathrm{M} = \mathrm{PL}. \end{cases}$$

*Then*

$$\Xi_{\alpha,\gamma} \le \min\left\{1, \varepsilon_{\alpha,\gamma}^2 + (1-\varepsilon_{\alpha,\gamma}^2)(\widetilde{\eta}_{\alpha,\gamma} + \bar{\gamma}_{\mathrm{cal}})\right\}. \tag{46}$$

*Moreover, the benchmark bias satisfies*

$$b_{\alpha,\gamma} \le \min\left\{H_{\mathrm{M}}, \frac{G_{\mathrm{M}}^2}{2\kappa_{\mathrm{M}}}\Xi_{\alpha,\gamma}\right\}, \tag{47}$$

*and the pessimistic approximation error satisfies*

$$\mathcal{E}_\Pi^{\mathrm{pess}} \le \frac{\Lambda_{\mathrm{M}}}{2}\mathcal{E}_\Pi + G_{\mathrm{M}}\sqrt{\Xi_{\alpha,\gamma}\mathcal{E}_\Pi} + \frac{G_{\mathrm{M}}^2}{2\kappa_{\mathrm{M}}}\Xi_{\alpha,\gamma}. \tag{48}$$

### D.6 Proof of the Main Upper Bounds

Having established the auxiliary results, we now prove Theorem 2 and Corollary 1.

**Proof of Theorem 2.** Throughout the proof, we condition on the pilot sample. Applying Lemma 7 with failure probability $\delta/2$, followed by (46), gives

$$\begin{aligned} \Xi_{\alpha,\gamma} &\le \min\left\{1,\ \varepsilon_{\alpha,\gamma}^2 + (1-\varepsilon_{\alpha,\gamma}^2)\left(\widetilde{\eta}_{\alpha,\gamma} + 2\gamma + \frac{4\log(2/\delta)}{3n_{\mathrm{cal}}}\right)\right\} \\ &\le 2\Psi_{\alpha,\gamma}(\delta), \end{aligned} \tag{49}$$

with probability at least $1-\delta/2$ over the calibration sample.

Conditional on the realized calibration sample, apply Lemma 5 with failure probability $\delta/2$ and the explicit localized radius in (40). The corresponding statistical term is

$$\xi_{n_{\mathrm{opt}}} = \frac{H_{\mathrm{M}}}{n_{\mathrm{opt}}} \max\left\{1024 C_{\mathrm{loc}}^2 d_{\Pi}, 64 \log(2/\delta)\right\}.$$

On this optimization-sample event, (56) and $\mathcal{E}_{\Pi}^{\mathrm{pess}} + \max\{\mathcal{E}_{\Pi}^{\mathrm{pess}}, \xi_{n_{\mathrm{opt}}}\} \le 2\mathcal{E}_{\Pi}^{\mathrm{pess}} + \xi_{n_{\mathrm{opt}}}$ imply

$$\mathbb{E}_D\left[\|x_{\widehat{\pi}_{\mathrm{LCP}}}(D) - x^\star(D)\|^2\right] \le \frac{4}{\kappa_{\mathrm{M}}} b_{\alpha,\gamma} + \frac{8}{\kappa_{\mathrm{M}}} \mathcal{E}_{\Pi}^{\mathrm{pess}} + \frac{4}{\kappa_{\mathrm{M}}} \xi_{n_{\mathrm{opt}}}. \tag{50}$$

Substituting the bias bounds from Lemma 8 and applying $2uv \le u^2 + v^2$ yields

$$\begin{aligned}
&\mathbb{E}_D\left[\|x_{\widehat{\pi}_{\mathrm{LCP}}}(D) - x^\star(D)\|^2\right] \\
\le& \frac{4\Lambda_{\mathrm{M}}}{\kappa_{\mathrm{M}}} \mathcal{E}_{\Pi} + \frac{8G_{\mathrm{M}}}{\kappa_{\mathrm{M}}} \sqrt{\Xi_{\alpha,\gamma} \mathcal{E}_{\Pi}} + \frac{6G_{\mathrm{M}}^2}{\kappa_{\mathrm{M}}^2} \Xi_{\alpha,\gamma} + \frac{4}{\kappa_{\mathrm{M}}} \xi_{n_{\mathrm{opt}}} \\
\le& \left(\frac{4\Lambda_{\mathrm{M}}}{\kappa_{\mathrm{M}}} + 4\right) \mathcal{E}_{\Pi} + \frac{10G_{\mathrm{M}}^2}{\kappa_{\mathrm{M}}^2} \Xi_{\alpha,\gamma} + \frac{4}{\kappa_{\mathrm{M}}} \xi_{n_{\mathrm{opt}}}.
\end{aligned}$$

By the tower property and a union bound, the calibration and optimization-sample events hold jointly with probability at least $1-\delta$, conditional on the pilot sample. On this joint event, (49) bounds the second term above by $20G_{\mathrm{M}}^2 \Psi_{\alpha,\gamma}(\delta)/\kappa_{\mathrm{M}}^2$. Moreover, since $C_{\mathrm{loc}} \ge 1$,

$$\frac{4}{\kappa_{\mathrm{M}}} \xi_{n_{\mathrm{opt}}} \le \frac{4096 H_{\mathrm{M}} C_{\mathrm{loc}}^2}{\kappa_{\mathrm{M}}} \frac{d_{\Pi} + \log(2/\delta)}{n_{\mathrm{opt}}}.$$

Thus, with

$$C_{\mathrm{M}} = \max\left\{\frac{4\Lambda_{\mathrm{M}}}{\kappa_{\mathrm{M}}} + 4, \frac{20G_{\mathrm{M}}^2}{\kappa_{\mathrm{M}}^2}, \frac{4096 H_{\mathrm{M}} C_{\mathrm{loc}}^2}{\kappa_{\mathrm{M}}}\right\}$$

we obtain

$$\mathbb{E}_D\left[\|x_{\widehat{\pi}_{\mathrm{LCP}}}(D) - x^\star(D)\|^2\right] \le C_{\mathrm{M}} \mathfrak{E}_{\alpha,\gamma}(\delta). \tag{51}$$

Substituting the BT and PL values of $\Lambda_{\mathrm{M}}$, $\kappa_{\mathrm{M}}$, $G_{\mathrm{M}}$, and $H_{\mathrm{M}}$ gives the respective constants in (27) and (28).

Finally, under the support and finiteness conditions of the theorem, (57) and (51) give

$$\begin{aligned}
\mathrm{SubOpt}_{\beta}(\widehat{\pi}_{\mathrm{LCP}}, \pi^\dagger) \le& \frac{C_{\mathrm{M}} C_{\mathrm{comp}}(\pi^\star; \Pi)}{2\beta} \mathfrak{E}_{\alpha,\gamma}(\delta) \\
&- \beta \mathbb{E}_{s\sim\rho}\left[D_{\mathrm{KL}}\left(\pi^\dagger(\cdot \mid s) \middle\| \pi^\star(\cdot \mid s)\right)\right],
\end{aligned}$$

which is (17). □

**Proof of Corollary 1.** Throughout, we condition on the pilot sample. Work on the joint calibration and optimization-sample event used in the proof of Theorem 2, which has conditional probability at least $1-\delta$. On this event, (49) holds, and (36) holds simultaneously for all $\pi \in \Pi$ with

$$\xi_{n_{\mathrm{opt}}} = \frac{H_{\mathrm{M}}}{n_{\mathrm{opt}}} \max\left\{1024 C_{\mathrm{loc}}^2 d_{\Pi}, 64 \log(2/\delta)\right\}.$$

For $\pi \in \Pi$, write

$$m_\pi = \mathbb{E}_D[e_\pi(D)], \qquad \widehat{m}_\pi = \frac{1}{n_{\mathrm{opt}}} \sum_{i=1}^{n_{\mathrm{opt}}} e_\pi(D_i^{\mathrm{opt}}).$$

Since

$$\widehat{L}(\pi) - \widehat{m}_\pi = \frac{1}{n_{\mathrm{opt}}} \sum_{i=1}^{n_{\mathrm{opt}}} f_{D_i^{\mathrm{opt}}}(\bar{x}(D_i^{\mathrm{opt}}))$$

is independent of $\pi$, we have $\widehat{m}_{\widetilde{\pi}} = \inf_{\pi\in\Pi} \widehat{m}_\pi + \Delta_{\mathrm{opt}}$. Moreover, (36) and $\max\{m_\pi, \xi_{n_{\mathrm{opt}}}\} \le m_\pi + \xi_{n_{\mathrm{opt}}}$ imply, simultaneously for all $\pi \in \Pi$,

$$m_\pi \le \frac{3}{2}\widehat{m}_\pi + \frac{1}{2}\xi_{n_{\mathrm{opt}}}, \qquad \widehat{m}_\pi \le \frac{4}{3}m_\pi + \frac{1}{3}\xi_{n_{\mathrm{opt}}}.$$

Therefore,

$$m_{\widetilde{\pi}} \le \frac{3}{2}\widehat{m}_{\widetilde{\pi}} + \frac{1}{2}\xi_{n_{\mathrm{opt}}} \le \frac{3}{2}\inf_{\pi\in\Pi}\widehat{m}_\pi + \frac{3}{2}\Delta_{\mathrm{opt}} + \frac{1}{2}\xi_{n_{\mathrm{opt}}} \le 2\mathcal{E}_\Pi^{\mathrm{pess}} + \xi_{n_{\mathrm{opt}}} + \frac{3}{2}\Delta_{\mathrm{opt}}.$$

The passage to the infimum does not require it to be attained.

Lemma 4 gives $\mathbb{E}_D[\|x_{\widetilde{\pi}}(D) - \bar{x}(D)\|^2] \le 2m_{\widetilde{\pi}}/\kappa_{\mathrm{M}}$. The benchmark-bias argument (55) in the proof of Theorem 6 gives $\mathbb{E}_D[\|\bar{x}(D) - x^\star(D)\|^2] \le 2b_{\alpha,\gamma}/\kappa_{\mathrm{M}}$. Applying the squared triangle inequality yields

$$\mathbb{E}_D\left[\|x_{\widetilde{\pi}}(D) - x^\star(D)\|^2\right] \le \frac{4}{\kappa_{\mathrm{M}}}\left(b_{\alpha,\gamma} + m_{\widetilde{\pi}}\right) \le \frac{4}{\kappa_{\mathrm{M}}}b_{\alpha,\gamma} + \frac{8}{\kappa_{\mathrm{M}}}\mathcal{E}_\Pi^{\mathrm{pess}} + \frac{4}{\kappa_{\mathrm{M}}}\xi_{n_{\mathrm{opt}}} + \frac{6}{\kappa_{\mathrm{M}}}\Delta_{\mathrm{opt}}.$$

By Lemma 8, (49), and the same constant calculation as in the proof of Theorem 2, the first three terms are bounded by $C_{\mathrm{M}}\mathfrak{E}_{\alpha,\gamma}(\delta)$. Thus,

$$\mathbb{E}_D\left[\|x_{\widetilde{\pi}}(D) - x^\star(D)\|^2\right] \le C_{\mathrm{M}}\mathfrak{E}_{\alpha,\gamma}(\delta) + \frac{6}{\kappa_{\mathrm{M}}}\Delta_{\mathrm{opt}}.$$

Finally, under the stated support and finiteness conditions, the KL representation of regularized suboptimality derived in the proof of Theorem 6 gives

$$\mathrm{SubOpt}_\beta(\widetilde{\pi}, \pi^\dagger) = \beta\mathbb{E}_{s\sim\rho}\left[D_{\mathrm{KL}}\left(\widetilde{\pi}(\cdot \mid s)\|\pi^\star(\cdot \mid s)\right)\right] - \beta\mathbb{E}_{s\sim\rho}\left[D_{\mathrm{KL}}\left(\pi^\dagger(\cdot \mid s)\middle\|\pi^\star(\cdot \mid s)\right)\right].$$

Applying Lemma 6 to the first term and then using the preceding score-error bound proves (29). The event is uniform over $\Pi$, so the conclusion holds simultaneously for every measurable, possibly data-dependent output and every comparator satisfying the stated conditions. □

### D.7 Proofs of Auxiliary Results

We provide the proofs of the auxiliary results below.

**Proof of Lemma 2.** Throughout the proof, we condition on the pilot and calibration samples and restrict all functions to $\mathcal{D}_0$. For a comparison-score function $v$, define

$$\|v\|_{\infty,2} = \sup_{D\in\mathcal{D}_0}\|v(D)\|,$$

and use $\|\cdot\|_\infty$ for the uniform norm of scalar-valued functions on $\mathcal{D}_0$. For a function class $\mathcal{F}$, let $\mathcal{N}(\eta, \mathcal{F}, \|\cdot\|)$ denote its $\eta$-covering number: the smallest number of balls of radius $\eta$, with centers in $\mathcal{F}$, needed to cover $\mathcal{F}$.

We first establish a uniform covering bound for the comparison-score class. Under the first condition, fix $\pi_0 \in \Pi$ and define

$$V = \mathrm{span}\{x_\pi - x_{\pi_0} : \pi \in \Pi\}.$$

Equipped with $\|\cdot\|_{\infty,2}$, this is a normed space of bounded functions with dimension at most $p$. The translated class $\{x_\pi - x_{\pi_0} : \pi \in \Pi\}$ is contained in the ball centered at zero with radius

$$R_{\mathrm{M}} = \begin{cases} 2B, & \mathrm{M} = \mathrm{BT}, \\ 2B\sqrt{K}, & \mathrm{M} = \mathrm{PL}. \end{cases}$$

If $\dim V = 0$, the score class is a singleton and its covering number is one. Otherwise, consider any finite collection of points in the translated class with pairwise distances greater than $\eta$. The balls of radius $\eta/2$ centered at these points are disjoint and contained in the ball of radius $R_{\mathrm{M}} + \eta/2$. Comparing volumes in $V$ bounds the cardinality of this collection by $(1+2R_{\mathrm{M}}/\eta)^{\dim V}$. A maximal such collection therefore exists and forms an $\eta$-cover. Since translation preserves distances and $\dim V \le p$, we obtain

$$\mathcal{N}\left(\eta, \{x_\pi : \pi \in \Pi\}, \|\cdot\|_{\infty,2}\right) \le \left(1 + \frac{A_x}{\eta}\right)^p, \qquad \eta > 0, \tag{52}$$

with $A_x = 2R_{\mathrm{M}}$.

Under the second condition, suppose first that $L_x > 0$. An internal $\eta/L_x$-net of $\mathcal{W}$ induces an $\eta$-net of the score class. Since $\mathcal{W}$ is contained in a Euclidean ball of radius $\mathrm{diam}(\mathcal{W})$, the same volumetric argument in $\mathbb{R}^p$ yields (52) with $A_x = 2L_x \,\mathrm{diam}(\mathcal{W})$. If $L_x = 0$, all comparison-score functions coincide, and the bound holds with $A_x = 0$.

We next transfer this covering bound to the excess-loss class. The outcome-wise negative log-likelihood is uniformly Lipschitz in the comparison score. Specifically,

$$|\ell_x(y, D) - \ell_z(y, D)| \le G_{\mathrm{M}} \|x - z\|,$$

where $G_{\mathrm{BT}} = 1$ and $G_{\mathrm{PL}} = \sqrt{2}(K-1)$. For BT, the derivative of the negative log-likelihood has absolute value at most one. For PL, each stagewise gradient is the difference between a probability vector over the remaining candidates and a standard basis vector, so its Euclidean norm is at most $\sqrt{2}$. Summing over the $K-1$ nontrivial stages gives the stated bound. Hence, for each fixed feasible $\boldsymbol{u}$, taking expectation under $P_{\mathrm{M},\boldsymbol{u}}(\cdot \mid D)$ preserves this Lipschitz bound. The uncertainty set does not depend on the candidate score. Taking suprema over the same set for $x$ and $z$, and applying the resulting inequality in both directions, therefore yields $|f_D(x) - f_D(z)| \le G_{\mathrm{M}} \|x - z\|$. Since the common benchmark term cancels in $e_\pi(D) - e_{\pi'}(D)$, it follows that

$$|e_\pi(D) - e_{\pi'}(D)| \le G_{\mathrm{M}} \|x_\pi(D) - x_{\pi'}(D)\|.$$

Consequently,

$$\|e_\pi - e_{\pi'}\|_\infty \le G_{\mathrm{M}} \|x_\pi - x_{\pi'}\|_{\infty,2}.$$

The bound $0 \le e_\pi \le H_{\mathrm{M}}$ implies that the mean constraint defining $\mathcal{H}_1$ is automatically satisfied. Thus,

$$\mathcal{H}_1 = \left\{ \frac{\lambda e_\pi}{H_{\mathrm{M}}} : \pi \in \Pi, \lambda \in [0,1] \right\}.$$

Moreover, for $\lambda, \lambda' \in [0,1]$,

$$\left\| \frac{\lambda e_\pi}{H_{\mathrm{M}}} - \frac{\lambda' e_{\pi'}}{H_{\mathrm{M}}} \right\|_\infty \le \frac{\lambda}{H_{\mathrm{M}}} \|e_\pi - e_{\pi'}\|_\infty + \frac{|\lambda - \lambda'|}{H_{\mathrm{M}}} \|e_{\pi'}\|_\infty \le \frac{G_{\mathrm{M}}}{H_{\mathrm{M}}} \|x_\pi - x_{\pi'}\|_{\infty,2} + |\lambda - \lambda'|. \quad (53)$$

Fix $0 < \eta \le 1$. By (52), the score class admits a cover of radius $H_{\mathrm{M}}\eta/(2G_{\mathrm{M}})$, with centers $x_{\pi_1}, \ldots, x_{\pi_{M_x}}$ and cardinality

$$M_x \le \left(1 + \frac{2G_{\mathrm{M}} A_x}{H_{\mathrm{M}} \eta}\right)^p.$$

For the scalar parameter, consider the grid

$$\lambda_k = \frac{k\eta}{2}, \qquad k = 0, \ldots, \left\lfloor \frac{2}{\eta} \right\rfloor.$$

All grid points belong to $[0,1]$, and every $\lambda \in [0,1]$ lies within $\eta/2$ of a grid point. The grid contains at most $1 + 2/\eta$ points.

Pairing the score representatives with these grid points gives the functions $\lambda_k e_{\pi_j}/H_{\mathrm{M}} \in \mathcal{H}_1$. For every $\lambda e_\pi / H_{\mathrm{M}} \in \mathcal{H}_1$, choose $j$ and $k$ from the respective covers. By (53),

$$\left\| \frac{\lambda e_\pi}{H_{\mathrm{M}}} - \frac{\lambda_k e_{\pi_j}}{H_{\mathrm{M}}} \right\|_\infty \le \frac{G_{\mathrm{M}}}{H_{\mathrm{M}}} \frac{H_{\mathrm{M}}\eta}{2G_{\mathrm{M}}} + \frac{\eta}{2} = \eta.$$

Consequently,

$$\mathcal{N}(\eta, \mathcal{H}_1, \|\cdot\|_\infty) \le \left(1 + \frac{2G_{\mathrm{M}} A_x}{H_{\mathrm{M}} \eta}\right)^p \left(1 + \frac{2}{\eta}\right).$$

Set

$$A = \max\left\{3, 1 + \frac{2G_{\mathrm{M}} A_x}{H_{\mathrm{M}}}\right\}.$$

Since $0 < \eta \le 1$ and $p \ge 1$, we obtain

$$\log \mathcal{N}(\eta, \mathcal{H}_1, \|\cdot\|_\infty) \le (p+1)\log(A/\eta) \le 2p \log(A/\eta).$$

The constant $A$ depends only on the fixed model and class bounds.

Fix $\zeta > 0$. Since $\mathcal{H}_\zeta \subseteq \mathcal{H}_1$, an $\eta/2$-cover of $\mathcal{H}_1$ induces an internal $\eta$-cover of $\mathcal{H}_\zeta$: discard balls that do not intersect $\mathcal{H}_\zeta$, and choose one member of $\mathcal{H}_\zeta$ from each remaining ball as a new center. The triangle inequality shows that doubling the radius preserves coverage. For $\eta_j = 2^{-j}$, $j \geq 1$, there exist finite $\eta_j$-nets $\mathcal{T}_j \subseteq \mathcal{H}_\zeta$ satisfying

$$\log|\mathcal{T}_j| \leq 2p \log\left(\frac{2A}{\eta_j}\right) = 2p\big[\log(2A) + j\log 2\big] \leq C_1 p(j+1),$$

where $C_1 = 2\max\{\log(2A), \log 2\}$. Set $\eta_0 = 1$ and $\mathcal{T}_0 = \{0\}$. Because $0 \in \mathcal{H}_\zeta$ and every $h \in \mathcal{H}_\zeta$ takes values in $[0,1]$, $\mathcal{T}_0$ is an internal $\eta_0$-net.

For any $h, g \in \mathcal{H}_\zeta$, nonnegativity and localization give

$$\mathbb{E}_D[(h(D) - g(D))^2] \leq \|h - g\|_\infty \mathbb{E}_D[h(D) + g(D)] \leq 2\zeta\|h - g\|_\infty.$$

Define the sets of nearby consecutive-net pairs by

$$\mathcal{A}_j = \{(h, g) \in \mathcal{T}_j \times \mathcal{T}_{j-1} : \|h - g\|_\infty \leq 3\eta_j\}.$$

These sets are nonempty, as can be seen by choosing consecutive-net approximants to the zero function. Moreover,

$$\log 2|\mathcal{A}_j| \leq \log 2 + \log|\mathcal{T}_j| + \log|\mathcal{T}_{j-1}| \leq \log 2 + C_1 p(j+1) + C_1 pj \leq C_2 p(j+1),$$

where $C_2 = 2C_1 + \log 2$. For $j = 1$, the same bound holds because $|\mathcal{T}_0| = 1$.

For every $(h, g) \in \mathcal{A}_j$, the random variables $\epsilon_i(h - g)(D_i^{\mathrm{opt}})$ are independent and centered, have absolute value at most $3\eta_j$, and satisfy

$$\mathbb{E}\left[\epsilon_i^2 (h - g)^2 (D_i^{\mathrm{opt}})\right] \leq 6\zeta\eta_j.$$

Bernstein's maximal inequality therefore gives

$$\begin{aligned}\mathbb{E} \max_{(h,g)\in\mathcal{A}_j} \left| \frac{1}{n_{\mathrm{opt}}} \sum_{i=1}^{n_{\mathrm{opt}}} \epsilon_i (h - g)(D_i^{\mathrm{opt}}) \right| &\leq \sqrt{\frac{12\zeta\eta_j \log(2|\mathcal{A}_j|)}{n_{\mathrm{opt}}}} + \frac{\eta_j \log(2|\mathcal{A}_j|)}{n_{\mathrm{opt}}} \\ &\leq C\left(\sqrt{\frac{p\zeta\eta_j(j+1)}{n_{\mathrm{opt}}}} + \frac{p\eta_j(j+1)}{n_{\mathrm{opt}}}\right).\end{aligned}$$

Here, $C$ depends only on $C_2$, and the expectation is over both the optimization sample and the Rademacher signs.

Finally, for each $h \in \mathcal{H}_\zeta$, choose $h^j \in \mathcal{T}_j$ with $\|h - h^j\|_\infty \leq \eta_j$, and set $h^0 = 0$. Then $(h^j, h^{j-1}) \in \mathcal{A}_j$. For every $J \geq 1$, the telescoping identity

$$h = (h - h^J) + \sum_{j=1}^{J} (h^j - h^{j-1})$$

and $|\epsilon_i| = 1$ imply that the remainder contributes at most $\eta_J$ to the absolute Rademacher average. Taking the supremum over $h \in \mathcal{H}_\zeta$ and then expectation yields

$$\mathfrak{R}_{n_{\mathrm{opt}}}(\zeta) \leq \eta_J + \sum_{j=1}^{J} \mathbb{E} \max_{(h,g)\in\mathcal{A}_j} \left| \frac{1}{n_{\mathrm{opt}}} \sum_{i=1}^{n_{\mathrm{opt}}} \epsilon_i (h - g)(D_i^{\mathrm{opt}}) \right|.$$

Substituting the preceding maximal bound and letting $J \to \infty$, we obtain

$$\begin{aligned}\mathfrak{R}_{n_{\mathrm{opt}}}(\zeta) &\leq C\sqrt{\frac{p\zeta}{n_{\mathrm{opt}}}} \sum_{j=1}^{\infty} \sqrt{2^{-j}(j+1)} + \frac{Cp}{n_{\mathrm{opt}}} \sum_{j=1}^{\infty} 2^{-j}(j+1) \\ &\leq C_{\mathrm{loc}}\left(\sqrt{\frac{p\zeta}{n_{\mathrm{opt}}}} + \frac{p}{n_{\mathrm{opt}}}\right).\end{aligned}$$

Both series converge. Since $C_1$ and $C_2$ depend only on the fixed model bounds and, in the second case, $L_x \operatorname{diam}(\mathcal{W})$, the resulting $C_{\mathrm{loc}}$ is independent of $p$, $n_{\mathrm{opt}}$, and $\zeta$. Increasing it if necessary ensures $C_{\mathrm{loc}} \geq 1$ and completes the proof. □

**Proof of Lemma 3.** Fix a target context $D$ for which the stated bounds hold, and let $x \in \mathcal{X}(D)$ and $y \in \mathcal{Y}(D)$. For BT feedback, $y \in \{0,1\}$ and $|x| \le B$ imply

$$\ell_x(y, D) = \log\left(1 + \exp(-(2y-1)x)\right) \le \log(1 + e^B).$$

For PL feedback, let $y = (y_1, \dots, y_K)$ be a complete ranking. Since the comparison scores are centered, $\sum_{k=1}^{K-1} x^{y_k} = -x^{y_K}$. Consequently,

$$\begin{aligned}\ell_x(y, D) &= \sum_{k=1}^{K-1}\left[\log\left(\sum_{\ell=k}^{K} e^{x^{y_\ell}}\right) - x^{y_k}\right] = \sum_{k=1}^{K-1}\log\left(\sum_{\ell=k}^{K} e^{x^{y_\ell}}\right) + x^{y_K} \\ &\le \sum_{k=1}^{K-1}\left[\log(K-k+1) + B\right] + B = \log(K!) + KB \le H_{\mathrm{PL}}.\end{aligned}$$

In both cases, $\ell_x(y, D) \ge 0$ because $p_{\mathrm{M},x}(y \mid D) \le 1$. Taking expectations under any $P_{\mathrm{M},\boldsymbol{u}}(\cdot \mid D)$ preserves these bounds. The empty-set convention ensures that $\widehat{\mathcal{U}}_\alpha(D;\gamma)$ is nonempty; hence taking the supremum over this set gives $0 \le f_D(x) \le H_{\mathrm{M}}$. □

**Proof of Lemma 4.** Fix a target context $D$ satisfying the stated assumptions. The argument in Lemma 3 applies to every $x \in \mathcal{X}(D)$ and yields $0 \le f_D(x) \le H_{\mathrm{M}}$ throughout this domain.

We first establish strong convexity. For BT feedback,

$$\ell_x(y, D) = -y\log\sigma(x) - (1-y)\log\sigma(-x) = \log(1 + e^x) - yx.$$

Consequently,

$$\frac{\partial^2}{\partial x^2}\ell_x(y, D) = \sigma(x)\sigma(-x) \ge \sigma(B)\sigma(-B) = \kappa_{\mathrm{BT}}, \qquad x \in [-B, B].$$

For PL feedback, let $\boldsymbol{\omega} = \mathrm{softmax}(x)$. The first-stage contribution to the Hessian is $\mathrm{diag}(\boldsymbol{\omega}) - \boldsymbol{\omega}\boldsymbol{\omega}^\top$. Each remaining stage is a log-sum-exp function minus a linear term and therefore has a positive semidefinite Hessian. Hence,

$$\nabla_x^2 \ell_x(y, D) \succeq \mathrm{diag}(\boldsymbol{\omega}) - \boldsymbol{\omega}\boldsymbol{\omega}^\top.$$

Moreover, since $\|x\|_\infty \le B$,

$$\omega^j = \frac{e^{x^j}}{\sum_{i=1}^K e^{x^i}} \ge \frac{e^{-B}}{\sum_{i=1}^K e^B} \ge \frac{e^{-2B}}{K}, \qquad j = 1, \dots, K.$$

Fix $\boldsymbol{h} \in \mathbf{1}_K^\perp$. Using $\sum_{j=1}^K \omega^j = 1$, we obtain, for any $c \in \mathbb{R}$,

$$\sum_{j=1}^K \omega^j (h^j - c)^2 = \sum_{j=1}^K \omega^j (h^j)^2 - \left(\sum_{j=1}^K \omega^j h^j\right)^2 + \left(c - \sum_{j=1}^K \omega^j h^j\right)^2.$$

Thus, the minimum over $c$ is attained at $c = \sum_{j=1}^K \omega^j h^j$, giving

$$\min_{c\in\mathbb{R}} \sum_{j=1}^K \omega^j (h^j - c)^2 = \sum_{j=1}^K \omega^j (h^j)^2 - \left(\sum_{j=1}^K \omega^j h^j\right)^2 = \boldsymbol{h}^\top\left[\mathrm{diag}(\boldsymbol{\omega}) - \boldsymbol{\omega}\boldsymbol{\omega}^\top\right]\boldsymbol{h}.$$

Similarly, since $\sum_{j=1}^K h^j = 0$,

$$\sum_{j=1}^K (h^j - c)^2 = \|\boldsymbol{h}\|^2 + Kc^2, \quad \Rightarrow \quad \min_{c\in\mathbb{R}} \sum_{j=1}^K (h^j - c)^2 = \|\boldsymbol{h}\|^2.$$

Combining these identities with the lower bound on $\omega^j$ yields

$$\begin{aligned}\boldsymbol{h}^\top \nabla_x^2 \ell_x(y, D)\boldsymbol{h} &\ge \boldsymbol{h}^\top\left[\mathrm{diag}(\boldsymbol{\omega}) - \boldsymbol{\omega}\boldsymbol{\omega}^\top\right]\boldsymbol{h} = \min_{c\in\mathbb{R}} \sum_{j=1}^K \omega^j (h^j - c)^2 \\ &\ge \frac{e^{-2B}}{K} \min_{c\in\mathbb{R}} \sum_{j=1}^K (h^j - c)^2 = \frac{e^{-2B}}{K}\|\boldsymbol{h}\|^2 = \kappa_{\mathrm{PL}}\|\boldsymbol{h}\|^2.\end{aligned}$$

The domain $\mathcal{X}(D)$ is convex, and in the PL case the difference between any two points in this domain belongs to $\mathbf{1}_K^\perp$. The preceding curvature bounds therefore establish that $\ell_x(y, D)$ is $\kappa_{\mathrm{M}}$-strongly convex on $\mathcal{X}(D)$ for any $\mathrm{M} \in \{\mathrm{BT}, \mathrm{PL}\}$ and every outcome $y$.

For fixed $D$, the latent uncertainty set and its indexed preference distributions do not depend on $x$. For each such distribution, taking the expectation preserves convexity of $x \mapsto \ell_x(y, D) - \kappa_{\mathrm{M}}\|x\|^2/2$. Taking the supremum over the same uncertainty set also preserves convexity. Consequently, $x \mapsto f_D(x) - \kappa_{\mathrm{M}}\|x\|^2/2$ is convex, so $f_D$ is $\kappa_{\mathrm{M}}$-strongly convex on $\mathcal{X}(D)$.

We next establish continuity. Because the suprema defining $f_D(x)$ and $f_D(z)$ range over the same set, for any $x, z \in \mathcal{X}(D)$,

$$\begin{aligned}|f_D(x) - f_D(z)| &\le \sup_{\boldsymbol{u} \in \widehat{\mathcal{U}}_\alpha(D;\gamma)} \mathbb{E}_{y \sim P_{\mathrm{M},\boldsymbol{u}}(\cdot|D)} \left[|\ell_x(y, D) - \ell_z(y, D)|\right] \\ &\le \max_{y \in \mathcal{Y}(D)} |\ell_x(y, D) - \ell_z(y, D)|.\end{aligned}$$

Since $\mathcal{Y}(D)$ is finite and each outcome-wise loss is continuous, the right-hand side tends to zero as $z \to x$. Thus, $f_D$ is continuous. Continuity and compactness of the nonempty domain $\mathcal{X}(D)$ guarantee that a minimizer exists, and strong convexity guarantees its uniqueness.

Finally, fix $x \in \mathcal{X}(D)$ and $t \in (0, 1)$. Convexity of $\mathcal{X}(D)$ ensures that $(1-t)\bar{x}(D) + tx$ remains in the same domain. By optimality of $\bar{x}(D)$ and strong convexity,

$$\begin{aligned}f_D(\bar{x}(D)) &\le f_D\big((1-t)\bar{x}(D) + tx\big) \\ &\le (1-t) f_D(\bar{x}(D)) + t f_D(x) - \frac{\kappa_{\mathrm{M}}}{2} t(1-t)\|x - \bar{x}(D)\|^2.\end{aligned}$$

Rearranging and dividing by $t > 0$ gives

$$f_D(x) - f_D(\bar{x}(D)) \ge \frac{\kappa_{\mathrm{M}}}{2}(1-t)\|x - \bar{x}(D)\|^2.$$

Letting $t \downarrow 0$ proves the quadratic lower bound in (33). The upper bound follows from $f_D(x) \le H_{\mathrm{M}}$ and $f_D(\bar{x}(D)) \ge 0$. □

**Proof of Lemma 5.** Throughout the proof, all probabilities and expectations are conditional on the pilot and calibration samples. Define

$$Z_{\zeta_{n_{\mathrm{opt}}}} = \sup_{h \in \mathcal{H}_{\zeta_{n_{\mathrm{opt}}}}} \left| \frac{1}{n_{\mathrm{opt}}} \sum_{i=1}^{n_{\mathrm{opt}}} h(D_i^{\mathrm{opt}}) - \mathbb{E}_D[h(D)] \right|.$$

We first bound its expectation by symmetrization. Let $D'_1, \ldots, D'_{n_{\mathrm{opt}}}$ be an independent copy of the optimization sample. Write $\mathbb{E}_{\mathrm{opt}}$, $\mathbb{E}'$, and $\mathbb{E}_\epsilon$ for expectations over the original sample, its independent copy, and the Rademacher signs, respectively, with the signs independent of both samples. For every $h \in \mathcal{H}_{\zeta_{n_{\mathrm{opt}}}}$,

$$\mathbb{E}_D[h(D)] = \mathbb{E}' \left[ \frac{1}{n_{\mathrm{opt}}} \sum_{i=1}^{n_{\mathrm{opt}}} h(D'_i) \right].$$

Thus, with all suprema in the following display taken over $h \in \mathcal{H}_{\zeta_{n_{\mathrm{opt}}}}$,

$$\begin{aligned}\mathbb{E}[Z_{\zeta_{n_{\mathrm{opt}}}}] &= \mathbb{E}_{\mathrm{opt}} \sup_{h \in \mathcal{H}_{\zeta_{n_{\mathrm{opt}}}}} \left| \mathbb{E}' \left[ \frac{1}{n_{\mathrm{opt}}} \sum_i \big(h(D_i^{\mathrm{opt}}) - h(D'_i)\big) \right] \right| \\ &\le \mathbb{E}_{\mathrm{opt}} \mathbb{E}' \sup_{h \in \mathcal{H}_{\zeta_{n_{\mathrm{opt}}}}} \left| \frac{1}{n_{\mathrm{opt}}} \sum_i \big(h(D_i^{\mathrm{opt}}) - h(D'_i)\big) \right| \\ &= \mathbb{E}_{\mathrm{opt}} \mathbb{E}' \mathbb{E}_\epsilon \sup_{h \in \mathcal{H}_{\zeta_{n_{\mathrm{opt}}}}} \left| \frac{1}{n_{\mathrm{opt}}} \sum_i \epsilon_i \big(h(D_i^{\mathrm{opt}}) - h(D'_i)\big) \right| \\ &\le \mathbb{E}_{\mathrm{opt}} \mathbb{E}_\epsilon \sup_{h \in \mathcal{H}_{\zeta_{n_{\mathrm{opt}}}}} \left| \frac{1}{n_{\mathrm{opt}}} \sum_i \epsilon_i h(D_i^{\mathrm{opt}}) \right| + \mathbb{E}' \mathbb{E}_\epsilon \sup_{h \in \mathcal{H}_{\zeta_{n_{\mathrm{opt}}}}} \left| \frac{1}{n_{\mathrm{opt}}} \sum_i \epsilon_i h(D'_i) \right| \\ &= 2\mathfrak{R}_{n_{\mathrm{opt}}}(\zeta_{n_{\mathrm{opt}}}).\end{aligned}$$

The first inequality follows from Jensen's inequality. The equality introducing the signs follows from invariance of the joint sample distribution under independent exchanges of the pairs $(D_i^{\mathrm{opt}}, D_i')$. The next inequality is the triangle inequality. The final line uses the identical distributions of the two samples and (34).

We next apply Bousquet's inequality (Bartlett et al., 2005, Theorem A.1) to the centered signed class

$$\left\{\pm\left(h(D) - \mathbb{E}_D[h(D)]\right) : h \in \mathcal{H}_{\zeta_{n_{\mathrm{opt}}}}\right\}.$$

Every member of this class has mean zero, is bounded in absolute value by one, and has variance at most $\zeta_{n_{\mathrm{opt}}}$. Conditional on the pilot and calibration samples, this class is fixed and the optimization contexts remain i.i.d. Since its supremum of empirical sums is $n_{\mathrm{opt}} Z_{\zeta_{n_{\mathrm{opt}}}}$, Bousquet's inequality, divided by $n_{\mathrm{opt}}$, gives, with conditional probability at least $1-\delta$,

$$\begin{aligned} Z_{\zeta_{n_{\mathrm{opt}}}} &\le \mathbb{E}[Z_{\zeta_{n_{\mathrm{opt}}}}] + \sqrt{\frac{2\big(\zeta_{n_{\mathrm{opt}}} + 2\mathbb{E}[Z_{\zeta_{n_{\mathrm{opt}}}}]\big)\log(1/\delta)}{n_{\mathrm{opt}}}} + \frac{\log(1/\delta)}{3n_{\mathrm{opt}}} \\ &\le \left(\frac{1}{8} + \sqrt{\frac{5}{128}} + \frac{1}{192}\right)\zeta_{n_{\mathrm{opt}}} < \frac{\zeta_{n_{\mathrm{opt}}}}{3}, \end{aligned}$$

where the second inequality uses the preceding expectation bound and (35).

Work on this event. For each $\pi \in \Pi$, set

$$\lambda_\pi = \frac{\xi_{n_{\mathrm{opt}}}}{\max\left\{\mathbb{E}_D[e_\pi(D)], \xi_{n_{\mathrm{opt}}}\right\}}.$$

Then $0 < \lambda_\pi \le 1$, and $\lambda_\pi \mathbb{E}_D[e_\pi(D)]/H_{\mathrm{M}} \le \zeta_{n_{\mathrm{opt}}}$. Hence, $\lambda_\pi e_\pi / H_{\mathrm{M}} \in \mathcal{H}_{\zeta_{n_{\mathrm{opt}}}}$. By the definition of $Z_{\zeta_{n_{\mathrm{opt}}}}$, simultaneously for all $\pi \in \Pi$,

$$\left|\frac{1}{n_{\mathrm{opt}}}\sum_{i=1}^{n_{\mathrm{opt}}} e_\pi(D_i^{\mathrm{opt}}) - \mathbb{E}_D[e_\pi(D)]\right| \le \frac{H_{\mathrm{M}} Z_{\zeta_{n_{\mathrm{opt}}}}}{\lambda_\pi} \le \frac{1}{3}\max\left\{\mathbb{E}_D[e_\pi(D)], \xi_{n_{\mathrm{opt}}}\right\}.$$

This proves (36).

Subtracting the policy-independent term $f_D(\bar{x}(D))$ does not affect empirical minimization. Thus, $\widehat{\pi}_{\mathrm{LCP}}$ also minimizes the empirical mean of $e_\pi$. For every $\pi \in \Pi$, the uniform deviation bound and empirical optimality therefore imply

$$\begin{aligned} \mathbb{E}_D[e_{\widehat{\pi}_{\mathrm{LCP}}}(D)] &\le \frac{1}{n_{\mathrm{opt}}}\sum_{i=1}^{n_{\mathrm{opt}}} e_{\widehat{\pi}_{\mathrm{LCP}}}(D_i^{\mathrm{opt}}) + \frac{1}{3}\max\left\{\mathbb{E}_D[e_{\widehat{\pi}_{\mathrm{LCP}}}(D)], \xi_{n_{\mathrm{opt}}}\right\} \\ &\le \frac{1}{n_{\mathrm{opt}}}\sum_{i=1}^{n_{\mathrm{opt}}} e_\pi(D_i^{\mathrm{opt}}) + \frac{1}{3}\max\left\{\mathbb{E}_D[e_{\widehat{\pi}_{\mathrm{LCP}}}(D)], \xi_{n_{\mathrm{opt}}}\right\} \\ &\le \mathbb{E}_D[e_\pi(D)] + \frac{1}{3}\max\left\{\mathbb{E}_D[e_\pi(D)], \xi_{n_{\mathrm{opt}}}\right\} + \frac{1}{3}\max\left\{\mathbb{E}_D[e_{\widehat{\pi}_{\mathrm{LCP}}}(D)], \xi_{n_{\mathrm{opt}}}\right\}. \end{aligned}$$

Taking the infimum over $\pi \in \Pi$ gives

$$\mathbb{E}_D[e_{\widehat{\pi}_{\mathrm{LCP}}}(D)] \le \mathcal{E}_\Pi^{\mathrm{pess}} + \frac{1}{3}\max\{\mathcal{E}_\Pi^{\mathrm{pess}}, \xi_{n_{\mathrm{opt}}}\} + \frac{1}{3}\max\{\mathbb{E}_D[e_{\widehat{\pi}_{\mathrm{LCP}}}(D)], \xi_{n_{\mathrm{opt}}}\}.$$

This step does not require the infimum to be attained, because $t \mapsto t + \max\{t, \xi_{n_{\mathrm{opt}}}\}/3$ is continuous and nondecreasing. Moreover, Lemma 4 gives $\mathbb{E}_D[e_{\widehat{\pi}_{\mathrm{LCP}}}(D)] \le H_{\mathrm{M}}$. If $\mathbb{E}_D[e_{\widehat{\pi}_{\mathrm{LCP}}}(D)] \le \max\{\mathcal{E}_\Pi^{\mathrm{pess}}, \xi_{n_{\mathrm{opt}}}\}$, then (37) follows immediately from $\mathcal{E}_\Pi^{\mathrm{pess}} \ge 0$. Otherwise, rearranging the preceding inequality yields

$$\mathbb{E}_D[e_{\widehat{\pi}_{\mathrm{LCP}}}(D)] \le \frac{3}{2}\mathcal{E}_\Pi^{\mathrm{pess}} + \frac{1}{2}\max\left\{\mathcal{E}_\Pi^{\mathrm{pess}}, \xi_{n_{\mathrm{opt}}}\right\} \le \mathcal{E}_\Pi^{\mathrm{pess}} + \max\left\{\mathcal{E}_\Pi^{\mathrm{pess}}, \xi_{n_{\mathrm{opt}}}\right\}.$$

Combining this with the bound by $H_{\mathrm{M}}$ proves (37).

Finally, Lemma 4 implies

$$\mathbb{E}_D\left[\|x_{\widehat{\pi}_{\mathrm{LCP}}}(D) - \bar{x}(D)\|^2\right] \le \frac{2\mathbb{E}_D[e_{\widehat{\pi}_{\mathrm{LCP}}}(D)]}{\kappa_{\mathrm{M}}}.$$

Applying (37) establishes (38). □

**Proof of Theorem 5.** Fix a target context $D$ for which the score bounds hold, and define

$$\begin{aligned} g_D(x) &= D_{\mathrm{KL}}\left(P_{\mathrm{M},x^\star(D)}(\cdot \mid D) \middle\| P_{\mathrm{M},x}(\cdot \mid D)\right) \\ &= \mathbb{E}_{y\sim P_{\mathrm{M},x^\star(D)}(\cdot|D)}\left[\ell_x(y,D) - \ell_{x^\star(D)}(y,D)\right]. \end{aligned}$$

We first establish global smoothness. Since the true preference distribution does not depend on $x$ and $\mathcal{Y}(D)$ is finite, differentiation commutes with the expectation defining $g_D$. For BT feedback,

$$0 \le g_D''(x) = \sigma(x)\sigma(-x) \le \frac{1}{4}.$$

For PL feedback, each stage contributes a positive semidefinite categorical covariance matrix to the Hessian of $\ell_x(y, D)$. For any probability vector $\boldsymbol{\omega}$, symmetry and the maximum absolute row-sum bound give

$$\left\|\mathrm{diag}(\boldsymbol{\omega}) - \boldsymbol{\omega}\boldsymbol{\omega}^\top\right\| \le \max_j 2\omega_j(1-\omega_j) \le \frac{1}{2}.$$

Embedding the stagewise matrices in the full score space, summing the $K-1$ contributions, and taking expectation under the true preference distribution therefore yields

$$0 \preceq \nabla^2 g_D(x) \preceq \frac{K-1}{2} I = \Lambda_{\mathrm{PL}} I.$$

Thus, in either model, $g_D$ is globally $\Lambda_{\mathrm{M}}$-smooth.

Because $g_D$ is nonnegative, applying the smoothness inequality at $z - \nabla g_D(z)/\Lambda_{\mathrm{M}}$ gives

$$0 \le g_D\left(z - \frac{\nabla g_D(z)}{\Lambda_{\mathrm{M}}}\right) \le g_D(z) - \frac{\|\nabla g_D(z)\|^2}{2\Lambda_{\mathrm{M}}}.$$

The auxiliary point $z - \nabla g_D(z)/\Lambda_{\mathrm{M}}$ need not belong to $\mathcal{X}(D)$: $g_D$ is nonnegative and globally $\Lambda_{\mathrm{M}}$-smooth on the unrestricted score space, with the PL likelihood defined by the same formula for all $x \in \mathbb{R}^K$. Consequently,

$$\|\nabla g_D(z)\| \le \sqrt{2\Lambda_{\mathrm{M}} g_D(z)}.$$

Applying smoothness once more and using Cauchy–Schwarz inequality, we obtain

$$g_D(x) \le g_D(z) + \langle \nabla g_D(z), x - z\rangle + \frac{\Lambda_{\mathrm{M}}}{2}\|x-z\|^2 \le \left(\sqrt{g_D(z)} + \sqrt{\frac{\Lambda_{\mathrm{M}}}{2}}\|x-z\|\right)^2.$$

Set $z = \bar{x}(D)$ and $x = x_{\widehat{\pi}_{\mathrm{LCP}}}(D)$. Applying Minkowski's inequality in $L^2$ over the target context distribution gives

$$\begin{aligned} \sqrt{\mathcal{R}(\widehat{\pi}_{\mathrm{LCP}})} &= \sqrt{\mathbb{E}_D[g_D(x_{\widehat{\pi}_{\mathrm{LCP}}}(D))]} \\ &\le \sqrt{\mathbb{E}_D\left[\left(\sqrt{g_D(\bar{x}(D))} + \sqrt{\frac{\Lambda_{\mathrm{M}}}{2}}\|x_{\widehat{\pi}_{\mathrm{LCP}}}(D) - \bar{x}(D)\|\right)^2\right]} \\ &\le \sqrt{\mathbb{E}_D[g_D(\bar{x}(D))]} + \sqrt{\frac{\Lambda_{\mathrm{M}}}{2}\mathbb{E}_D\left[\|x_{\widehat{\pi}_{\mathrm{LCP}}}(D) - \bar{x}(D)\|^2\right]}. \end{aligned}$$

The first term is $\sqrt{b_{\alpha,\gamma}}$. Substituting (38) into the second term proves (39) on the event of Lemma 5.

Finally, for every $\pi \in \Pi$, nonnegativity of $\ell_{x^\star(D)}(y, D)$ and Lemma 3 imply

$$g_D(x_\pi(D)) \le \mathbb{E}_{y\sim P_{\mathrm{M},x^\star(D)}(\cdot|D)}[\ell_{x_\pi(D)}(y,D)] \le H_{\mathrm{M}}.$$

Taking expectation over $D$ proves $\mathcal{R}(\pi) \le H_{\mathrm{M}}$. □

**Proof of Corollary 2.** The choice in (40) gives

$$\frac{d_\Pi}{n_{\mathrm{opt}}} \le \frac{\zeta_{n_{\mathrm{opt}}}}{1024 C_{\mathrm{loc}}^2}.$$

Hence, using $C_{\mathrm{loc}} \geq 1$,

$$16C_{\mathrm{loc}}\sqrt{\frac{d_\Pi \zeta_{n_{\mathrm{opt}}}}{n_{\mathrm{opt}}}} \leq \frac{\zeta_{n_{\mathrm{opt}}}}{2}, \qquad \frac{16C_{\mathrm{loc}}d_\Pi}{n_{\mathrm{opt}}} \leq \frac{\zeta_{n_{\mathrm{opt}}}}{64C_{\mathrm{loc}}} \leq \frac{\zeta_{n_{\mathrm{opt}}}}{2}.$$

Combining these inequalities with (26) gives $16\mathfrak{R}_{n_{\mathrm{opt}}}(\zeta_{n_{\mathrm{opt}}}) \leq \zeta_{n_{\mathrm{opt}}}$. The remaining condition in (35) holds by construction.

We may therefore apply Theorem 5 with $\xi_{n_{\mathrm{opt}}} = H_{\mathrm{M}}\zeta_{n_{\mathrm{opt}}}$. Using the inequalities $\min\{H_{\mathrm{M}}, A\} \leq A$, $(\sqrt{u}+\sqrt{v})^2 \leq 2u+2v$ and

$$\mathcal{E}_\Pi^{\mathrm{pess}} + \max\left\{\mathcal{E}_\Pi^{\mathrm{pess}}, \xi_{n_{\mathrm{opt}}}\right\} \leq 2\mathcal{E}_\Pi^{\mathrm{pess}} + \xi_{n_{\mathrm{opt}}},$$

we obtain (41). The additional cap follows from $\mathcal{R}(\widehat{\pi}_{\mathrm{LCP}}) \leq H_{\mathrm{M}}$. □

**Proof of Lemma 6.** Fix a policy $\pi$ satisfying the stated conditions. Let $\mu$ denote the joint distribution of $(s, a, a')$ obtained by drawing $s \sim \rho$ and, conditionally on $s$, drawing $a \sim \pi^\star(\cdot \mid s)$ and $a' \sim \pi(\cdot \mid s)$ independently. Let $\widetilde{\mu}$ be the distribution obtained by swapping the two responses. The common-support condition implies that these measures are mutually absolutely continuous.

Define

$$T(s,a,a') = \log\frac{\mathrm{d}\mu}{\mathrm{d}\widetilde{\mu}}(s,a,a') = \log\frac{\pi^\star(a \mid s)\pi(a' \mid s)}{\pi(a \mid s)\pi^\star(a' \mid s)}.$$

By (2) and the definition of $r_\pi$, the prompt-dependent normalizing term cancels, giving

$$\beta T(s,a,a') = [r_\pi(s,a') - r^\star(s,a')] - [r_\pi(s,a) - r^\star(s,a)].$$

Consequently, the definition of $C_{\mathrm{comp}}(\pi^\star;\Pi)$ yields

$$\begin{aligned}\mathbb{E}_\mu[T^2] &= \frac{1}{\beta^2}\mathbb{E}_{\substack{s\sim\rho \\ a\sim\pi^\star(\cdot|s), a'\sim\pi(\cdot|s)}}\left[\left([r_\pi(s,a) - r^\star(s,a)] - [r_\pi(s,a') - r^\star(s,a')]\right)^2\right] \\ &\leq \frac{C_{\mathrm{comp}}(\pi^\star;\Pi)}{\beta^2}\mathbb{E}_D\left[\|x_\pi(D) - x^\star(D)\|^2\right] < \infty,\end{aligned} \tag{54}$$

where finiteness follows from the bounded comparison scores and the assumed finite comparison coefficient.

Swapping $a$ and $a'$ exchanges $\mu$ and $\widetilde{\mu}$ and changes $T$ to $-T$. Therefore,

$$\mathbb{E}_{\widetilde{\mu}}[T^2] = \mathbb{E}_\mu[T^2], \qquad \mathbb{E}_{\widetilde{\mu}}[T] = -\mathbb{E}_\mu[T].$$

The first moments are finite by Cauchy–Schwarz. Using $\mathrm{d}\mu = e^T\mathrm{d}\widetilde{\mu}$ and $t\tanh(t/2) \leq t^2/2$ for every $t \in \mathbb{R}$, we obtain

$$\begin{aligned}D_{\mathrm{KL}}(\mu\|\widetilde{\mu}) = \mathbb{E}_\mu[T] = \frac{1}{2}\int T(\mathrm{d}\mu - \mathrm{d}\widetilde{\mu}) &\overset{\text{(i)}}{=} \frac{1}{2}\int T\tanh(T/2)(\mathrm{d}\mu + \mathrm{d}\widetilde{\mu}) \\ &\leq \frac{1}{4}\int T^2(\mathrm{d}\mu + \mathrm{d}\widetilde{\mu}) = \frac{1}{2}\mathbb{E}_\mu[T^2],\end{aligned}$$

where equality (i) follows from

$$\mathrm{d}\mu - \mathrm{d}\widetilde{\mu} = (e^T - 1)\mathrm{d}\widetilde{\mu} = \frac{e^T-1}{e^T+1}(e^T+1)\mathrm{d}\widetilde{\mu} = \frac{e^T-1}{e^T+1}(\mathrm{d}\mu + \mathrm{d}\widetilde{\mu}).$$

Both $\mu$ and $\widetilde{\mu}$ have prompt marginal $\rho$. Hence, the prompt-marginal term in the KL chain rule vanishes, since $D_{\mathrm{KL}}(\rho\|\rho) = 0$. Conditional on $s$, their joint response distributions are $\pi^\star(\cdot \mid s) \otimes \pi(\cdot \mid s)$ and $\pi(\cdot \mid s) \otimes \pi^\star(\cdot \mid s)$, respectively. Applying the chain rule and expanding the conditional log-likelihood ratio therefore gives

$$\begin{aligned}D_{\mathrm{KL}}(\mu\|\widetilde{\mu}) &= \mathbb{E}_{s\sim\rho}\mathbb{E}_{\substack{a\sim\pi^\star(\cdot|s) \\ a'\sim\pi(\cdot|s)}}\left[\log\frac{\pi^\star(a \mid s)\pi(a' \mid s)}{\pi(a \mid s)\pi^\star(a' \mid s)}\right] \\ &= \mathbb{E}_{s\sim\rho}\left[D_{\mathrm{KL}}\left(\pi^\star(\cdot \mid s)\|\pi(\cdot \mid s)\right) + D_{\mathrm{KL}}\left(\pi(\cdot \mid s)\|\pi^\star(\cdot \mid s)\right)\right].\end{aligned}$$

Dropping the nonnegative $D_{\mathrm{KL}}(\pi^\star(\cdot \mid s)\|\pi(\cdot \mid s))$ term and applying (54) proves (42). □

**Proof of Theorem 6.** Work on the event of Lemma 5, which has conditional probability at least $1-\delta$. Consider a realization for which $\widehat{\pi}_{\mathrm{LCP}}$ shares the response support of $\pi^\star$ for $\rho$-almost every prompt $s$.

We first bound the comparison-score error relative to the true score. For almost every target context $D$, recall that

$$\begin{aligned} g_D(x) &= D_{\mathrm{KL}}\left(P_{\mathrm{M},x^\star(D)}(\cdot \mid D)\middle\| P_{\mathrm{M},x}(\cdot \mid D)\right) \\ &= \mathbb{E}_{y\sim P_{\mathrm{M},x^\star(D)}(\cdot|D)}\left[\ell_x(y,D)-\ell_{x^\star(D)}(y,D)\right]. \end{aligned}$$

By the proof of Lemma 4, each outcome-wise loss $\ell_x(y,D)$ is $\kappa_{\mathrm{M}}$-strongly convex on $\mathcal{X}(D)$. Taking expectation under a distribution independent of $x$ and subtracting an $x$-independent term preserve the same strong-convexity constant. Hence, $g_D$ is also $\kappa_{\mathrm{M}}$-strongly convex on $\mathcal{X}(D)$. Since $g_D(x)\geq 0$ and $g_D(x^\star(D))=0$, the true comparison score $x^\star(D)$ minimizes $g_D$ over $\mathcal{X}(D)$. Both $x^\star(D)$ and $x_\pi(D)$ belong to this domain, so quadratic growth gives

$$g_D(x_\pi(D)) \geq \frac{\kappa_{\mathrm{M}}}{2}\|x_\pi(D)-x^\star(D)\|^2.$$

Since $\bar{x}(D)\in\mathcal{X}(D)$, applying this bound at $x=\bar{x}(D)$ and taking expectation over $D$ gives

$$\mathbb{E}_D\left[\|\bar{x}(D)-x^\star(D)\|^2\right] \leq \frac{2}{\kappa_{\mathrm{M}}}\mathbb{E}_D[g_D(\bar{x}(D))] = \frac{2b_{\alpha,\gamma}}{\kappa_{\mathrm{M}}}. \tag{55}$$

Combining this inequality with (38) and applying the triangle inequality in $L^2$ yields

$$\begin{aligned} &\sqrt{\mathbb{E}_D\left[\|x_{\widehat{\pi}_{\mathrm{LCP}}}(D)-x^\star(D)\|^2\right]} \\ \leq &\sqrt{\mathbb{E}_D\left[\|x_{\widehat{\pi}_{\mathrm{LCP}}}(D)-\bar{x}(D)\|^2\right]}+\sqrt{\mathbb{E}_D\left[\|\bar{x}(D)-x^\star(D)\|^2\right]} \\ \leq &\sqrt{\frac{2}{\kappa_{\mathrm{M}}}}\left(\sqrt{b_{\alpha,\gamma}}+\sqrt{\mathcal{E}_\Pi^{\mathrm{pess}}+\max\left\{\mathcal{E}_\Pi^{\mathrm{pess}},\xi_{n_{\mathrm{opt}}}\right\}}\right). \end{aligned}$$

Independently, the same curvature argument and the deterministic risk bound in Theorem 5 give

$$\mathbb{E}_D\left[\|x_{\widehat{\pi}_{\mathrm{LCP}}}(D)-x^\star(D)\|^2\right] \leq \frac{2}{\kappa_{\mathrm{M}}}\mathcal{R}(\widehat{\pi}_{\mathrm{LCP}}) \leq \frac{2H_{\mathrm{M}}}{\kappa_{\mathrm{M}}}.$$

Consequently,

$$\mathbb{E}_D\left[\|x_{\widehat{\pi}_{\mathrm{LCP}}}(D)-x^\star(D)\|^2\right] \leq \frac{2}{\kappa_{\mathrm{M}}}\min\left\{H_{\mathrm{M}},\left(\sqrt{b_{\alpha,\gamma}}+\sqrt{\mathcal{E}_\Pi^{\mathrm{pess}}+\max\left\{\mathcal{E}_\Pi^{\mathrm{pess}},\xi_{n_{\mathrm{opt}}}\right\}}\right)^2\right\}. \tag{56}$$

We now establish the Gibbs identity for the regularized objective. By (2), for each policy $\pi$ under consideration, the following identity holds for $\rho$-almost every $s$ and $\pi(\cdot\mid s)$-almost every $a$:

$$r^\star(s,a)-\beta\log\frac{\pi(a\mid s)}{\pi_{\mathrm{ref}}(a\mid s)} = \beta\log Z_\beta^\star(s)-\beta\log\frac{\pi(a\mid s)}{\pi^\star(a\mid s)}.$$

The assumed finiteness of $J$ and $H$ justifies the following expectations. In particular, finite $H(\pi)$ implies integrability of the policy log-ratio, since its negative part has expectation at most $1/e$. Applying the preceding identity first to $\pi^\star$ therefore shows that $\log Z_\beta^\star(s)$ is integrable. Taking expectation under $s\sim\rho$ and $a\sim\pi(\cdot\mid s)$ yields

$$J_\beta(\pi) = \beta\mathbb{E}_{s\sim\rho}\log Z_\beta^\star(s)-\beta\mathbb{E}_{s\sim\rho}\left[D_{\mathrm{KL}}\left(\pi(\cdot\mid s)\|\pi^\star(\cdot\mid s)\right)\right].$$

Setting $\pi=\pi^\star$ gives $J_\beta(\pi^\star)=\beta\mathbb{E}_{s\sim\rho}\log Z_\beta^\star(s)$. Hence,

$$J_\beta(\pi^\star)-J_\beta(\pi) = \beta\mathbb{E}_{s\sim\rho}\left[D_{\mathrm{KL}}\left(\pi(\cdot\mid s)\|\pi^\star(\cdot\mid s)\right)\right].$$

Applying this identity to $\pi=\widehat{\pi}_{\mathrm{LCP}}$ and $\pi=\pi^\dagger$ and subtracting gives

$$\begin{aligned} \mathrm{SubOpt}_\beta(\widehat{\pi}_{\mathrm{LCP}},\pi^\dagger) &= J_\beta(\pi^\dagger)-J_\beta(\widehat{\pi}_{\mathrm{LCP}}) = J_\beta(\pi^\star)-J_\beta(\widehat{\pi}_{\mathrm{LCP}})-\left(J_\beta(\pi^\star)-J_\beta(\pi^\dagger)\right) \\ &= \beta\mathbb{E}_{s\sim\rho}\left[D_{\mathrm{KL}}\left(\widehat{\pi}_{\mathrm{LCP}}(\cdot\mid s)\|\pi^\star(\cdot\mid s)\right)\right]-\beta\mathbb{E}_{s\sim\rho}\left[D_{\mathrm{KL}}\left(\pi^\dagger(\cdot\mid s)\middle\|\pi^\star(\cdot\mid s)\right)\right] \end{aligned}$$

Applying Lemma 6 to the learned policy yields

$$\begin{aligned} \mathrm{SubOpt}_\beta(\widehat{\pi}_{\mathrm{LCP}},\pi^\dagger) \leq &\frac{C_{\mathrm{comp}}(\pi^\star;\Pi)}{2\beta}\mathbb{E}_D\left[\|x_{\widehat{\pi}_{\mathrm{LCP}}}(D)-x^\star(D)\|^2\right] \\ &-\beta\mathbb{E}_{s\sim\rho}\left[D_{\mathrm{KL}}\left(\pi^\dagger(\cdot\mid s)\middle\|\pi^\star(\cdot\mid s)\right)\right]. \end{aligned} \tag{57}$$

Substituting (56) proves (43). □

**Proof of Lemma 7.** Throughout the proof, we condition on the pilot sample. Let $F_S$ denote the cumulative distribution function of the target nonconformity score, with $F_S(+\infty) = 1$. For each realized calibration sample, model correctness and the definition of the conformal prediction set imply

$$\bar{\gamma}_{\mathrm{cal}} = 1 - \mathbb{P}_{(D,y)}\left\{\widehat{s}(D,y) \le \widehat{q}^{\mathrm{cal}}_{1-\gamma}\right\} = 1 - F_S\big(\widehat{q}^{\mathrm{cal}}_{1-\gamma}\big).$$

Recall $k_\gamma = \lceil (n_{\mathrm{cal}}+1)(1-\gamma) \rceil$. If $k_\gamma = n_{\mathrm{cal}} + 1$, then $\widehat{q}^{\mathrm{cal}}_{1-\gamma} = +\infty$ and $\bar{\gamma}_{\mathrm{cal}} = 0$. It therefore suffices to consider $k_\gamma \le n_{\mathrm{cal}}$. Set

$$\tau = 2\gamma + \frac{4\log(1/\delta_{\mathrm{cal}})}{3n_{\mathrm{cal}}}.$$

The claim is immediate if $\tau \ge 1$. Otherwise, define the population quantile

$$q_\tau = \inf\left\{t \in [0,+\infty] : F_S(t) \ge 1-\tau\right\},$$

allowing $q_\tau = +\infty$. Its defining properties give $F_S(q_\tau-) \le 1-\tau \le F_S(q_\tau)$, and hence

$$p_\tau = \mathbb{P}\{S_i^{\mathrm{cal}} \ge q_\tau\} = 1 - F_S(q_\tau-) \ge \tau.$$

Here, $F_S(q_\tau-)$ denotes the left limit.

On the event $\{\bar{\gamma}_{\mathrm{cal}} > \tau\}$, we have $F_S(\widehat{q}^{\mathrm{cal}}_{1-\gamma}) < 1-\tau$, and therefore $\widehat{q}^{\mathrm{cal}}_{1-\gamma} < q_\tau$. Since the calibrated threshold is the $k_\gamma$-th order statistic, at least $k_\gamma$ calibration scores are strictly below $q_\tau$. Consequently, on this event,

$$N_\tau = \sum_{i \in \mathcal{I}_{\mathrm{cal}}} \mathbb{I}\{S_i^{\mathrm{cal}} \ge q_\tau\} \le n_{\mathrm{cal}} - k_\gamma < n_{\mathrm{cal}}\gamma.$$

By the conditional i.i.d. assumption, $N_\tau \sim \mathrm{Binomial}(n_{\mathrm{cal}}, p_\tau)$. Applying Markov's inequality to $4^{-N_\tau}$ and using $1-u \le e^{-u}$ yield

$$\mathbb{P}\{\bar{\gamma}_{\mathrm{cal}} > \tau\} \le \mathbb{P}\left\{N_\tau \le n_{\mathrm{cal}}\gamma\right\} \le 4^{n_{\mathrm{cal}}\gamma}\mathbb{E}[4^{-N_\tau}] = 4^{n_{\mathrm{cal}}\gamma}\left(1 - \frac{3p_\tau}{4}\right)^{n_{\mathrm{cal}}}$$

$$\le \exp\left(n_{\mathrm{cal}}\gamma\log 4 - \frac{3n_{\mathrm{cal}}p_\tau}{4}\right) \le \exp\left(n_{\mathrm{cal}}\gamma\log 4 - \frac{3n_{\mathrm{cal}}\tau}{4}\right) \le \delta_{\mathrm{cal}}.$$

Thus, conditional on the pilot sample, with probability at least $1-\delta_{\mathrm{cal}}$,

$$\bar{\gamma}_{\mathrm{cal}} \le \min\left\{1, 2\gamma + \frac{4\log(1/\delta_{\mathrm{cal}})}{3n_{\mathrm{cal}}}\right\}.$$

To prove the sharper statement, note that $k_\gamma = n_{\mathrm{cal}} + 1 - j_\gamma$. The case $j_\gamma = 0$ was already treated. For $j_\gamma \ge 1$, introduce i.i.d. $V_i \sim \mathrm{Unif}(0,1)$, independent of the calibration scores, solely for the proof, and set

$$W_i = F_S(S_i^{\mathrm{cal}}-) + V_i\big[F_S(S_i^{\mathrm{cal}}) - F_S(S_i^{\mathrm{cal}}-)\big].$$

The randomized probability integral transform gives i.i.d. uniform $W_i$, including when the scores have atoms at $+\infty$. Since $W_i \le F_S(S_i^{\mathrm{cal}})$, their order statistics satisfy $W_{(k_\gamma)} \le F_S(S^{\mathrm{cal}}_{(k_\gamma)})$. Consequently,

$$\bar{\gamma}_{\mathrm{cal}} \le 1 - W_{(k_\gamma)}, \qquad 1 - W_{(k_\gamma)} \sim \mathrm{Beta}(j_\gamma, n_{\mathrm{cal}} + 1 - j_\gamma).$$

By the definition of $u_{n_{\mathrm{cal}},\gamma}(\delta_{\mathrm{cal}})$,

$$\mathbb{P}\left\{\bar{\gamma}_{\mathrm{cal}} \le u_{n_{\mathrm{cal}},\gamma}(\delta_{\mathrm{cal}})\right\} \ge \mathbb{P}\left\{1 - W_{(k_\gamma)} \le u_{n_{\mathrm{cal}},\gamma}(\delta_{\mathrm{cal}})\right\} = 1 - \delta_{\mathrm{cal}}.$$

This proves the first inequality in (44). If the conditional score distribution is continuous, then $\bar{\gamma}_{\mathrm{cal}} = 1 - W_{(k_\gamma)}$ almost surely, so the beta distribution is exact.

If $\tau < 1$, define

$$\widetilde{N}_\tau = \sum_{i \in \mathcal{I}_{\mathrm{cal}}} \mathbb{I}\{W_i \ge 1-\tau\} \sim \mathrm{Binomial}(n_{\mathrm{cal}}, \tau).$$

On the event $\{1 - W_{(k_\gamma)} > \tau\}$, $\widetilde{N}_\tau \le n_{\mathrm{cal}} - k_\gamma < n_{\mathrm{cal}}\gamma$. The same exponential-moment bound as above therefore gives

$$\mathbb{P}\{1 - W_{(k_\gamma)} > \tau\} \le 4^{n_{\mathrm{cal}}\gamma}(1 - 3\tau/4)^{n_{\mathrm{cal}}} \le \delta_{\mathrm{cal}}.$$

Hence $u_{n_{\mathrm{cal}},\gamma}(\delta_{\mathrm{cal}}) \le \tau$. If $\tau \ge 1$, the bound $u_{n_{\mathrm{cal}},\gamma}(\delta_{\mathrm{cal}}) \le 1$ is immediate. Together, these cases prove the second inequality in (44). □

**Proof of Lemma 8.** Define

$$v_D = \omega_D + (1-\omega_D)\max\{\varepsilon_{\alpha,\gamma}, t_D\}.$$

Then $0 \le v_D \le 1$ and $\Xi_{\alpha,\gamma} = \mathbb{E}_D[v_D^2]$. For $w, t \in [0,1]$, convexity of $t \mapsto [w+(1-w)t]^2$ on $[\varepsilon_{\alpha,\gamma}, 1]$ gives

$$[w+(1-w)\max\{\varepsilon_{\alpha,\gamma}, t\}]^2 \le a_\varepsilon(w) + (1-a_\varepsilon(w))t,$$

where $a_\varepsilon(w) = [w+(1-w)\varepsilon_{\alpha,\gamma}]^2 = \varepsilon_{\alpha,\gamma}^2 + (1-\varepsilon_{\alpha,\gamma}^2)r_\varepsilon(w)$. Indeed, for $t > \varepsilon_{\alpha,\gamma}$, the chord coefficient $(t-\varepsilon_{\alpha,\gamma})/(1-\varepsilon_{\alpha,\gamma})$ is at most $t$; for $t \le \varepsilon_{\alpha,\gamma}$, the left-hand side equals $a_\varepsilon(w)$. Taking expectations therefore yields

$$\begin{aligned}\Xi_{\alpha,\gamma} &\le \varepsilon_{\alpha,\gamma}^2 + (1-\varepsilon_{\alpha,\gamma}^2)\mathbb{E}_D\left[r_\varepsilon(\omega_D) + (1-r_\varepsilon(\omega_D))t_D\right] \\ &\le \varepsilon_{\alpha,\gamma}^2 + (1-\varepsilon_{\alpha,\gamma}^2)(\widetilde{\eta}_{\alpha,\gamma} + \bar{\gamma}_{\mathrm{cal}}).\end{aligned}$$

Together with $v_D \le 1$, this proves (46).

Fix a context $D$ for which the score bounds hold. We first show that $G_{\mathrm{M}}$ bounds the diameter of the outcome-wise gradients, i.e.,

$$\sup_{y,y' \in \mathcal{Y}(D)} \|\nabla_x \ell_x(y,D) - \nabla_x \ell_x(y',D)\| \le G_{\mathrm{M}}.$$

For BT feedback, this follows directly from $\partial_x \ell_x(y,D) = \sigma(x) - y$. For PL feedback, let $\boldsymbol{\omega}_k(x,y)$ denote the stage-$k$ choice-probability vector, with zero coordinates for previously selected candidates. Including the deterministic final stage gives

$$\nabla_x \ell_x(y,D) = \sum_{k=1}^{K} \boldsymbol{\omega}_k(x,y) - \mathbf{1}_K.$$

Since $\boldsymbol{\omega}_1(x,y) = \mathrm{softmax}(x)$ is independent of the ranking, the first-stage vectors cancel when subtracting the gradients. Moreover, any two probability vectors have Euclidean distance at most $\sqrt{2}$. The triangle inequality therefore yields

$$\begin{aligned}\|\nabla_x \ell_x(y,D) - \nabla_x \ell_x(y',D)\| &= \left\| \sum_{k=2}^{K} (\boldsymbol{\omega}_k(x,y) - \boldsymbol{\omega}_k(x,y')) \right\| \\ &\le \sum_{k=2}^{K} \|\boldsymbol{\omega}_k(x,y) - \boldsymbol{\omega}_k(x,y')\| \le \sqrt{2}(K-1) \le G_{\mathrm{PL}}.\end{aligned}$$

If $R_k(y) = R_k(y')$, the corresponding PL choice-probability vectors also coincide. Applying the same argument only at stages where these sets differ, and using the BT gradient formula for binary feedback, gives, whenever $\widehat{\mathcal{C}}_\gamma(D) \ne \varnothing$,

$$\sup_{x \in \mathcal{X}(D)} \max_{y,y' \in \widehat{\mathcal{C}}_\gamma(D)} \|\nabla_x \ell_x(y,D) - \nabla_x \ell_x(y',D)\| \le G_{\mathrm{M}} \omega_D. \tag{58}$$

Fix any $\boldsymbol{u} \in \widehat{\mathcal{U}}_\alpha(D;\gamma)$ and write $Q = P_{\mathrm{M},\boldsymbol{u}}(\cdot \mid D)$ and $P_D^\star = P_{\mathrm{M},x^\star(D)}(\cdot \mid D)$. Recall that $\nabla g_D(x) = \mathbb{E}_{P_D^\star}[\nabla_x \ell_x(y,D)]$. In both models, any single outcome can receive probability arbitrarily close to one under latent scores in $\Theta(D)$. Thus, for nonempty $\widehat{\mathcal{C}}_\gamma(D)$, the compatibility constraint is feasible and gives $Q(\widehat{\mathcal{C}}_\gamma(D)) \ge 1-\varepsilon_{\alpha,\gamma}$; also $P_D^\star(\widehat{\mathcal{C}}_\gamma(D)) = 1 - t_D$. Couple $Y \sim Q$ and $Y' \sim P_D^\star$ so that

$$\mathbb{P}\{Y \in \widehat{\mathcal{C}}_\gamma(D),\ Y' \in \widehat{\mathcal{C}}_\gamma(D)\} = \min\{Q(\widehat{\mathcal{C}}_\gamma(D)), P_D^\star(\widehat{\mathcal{C}}_\gamma(D))\} \ge 1 - \max\{\varepsilon_{\alpha,\gamma}, t_D\}.$$

Such a coupling is obtained by matching the two membership indicators with maximal overlap and then sampling from the corresponding conditional distributions. On this event use (58), and on its complement use the global gradient-diameter bound. Since $0 \le \omega_D \le 1$, the triangle inequality gives

$$\begin{aligned}\|\mathbb{E}_Q[\nabla_x \ell_x(y,D)] - \nabla g_D(x)\| &\le \mathbb{E}\|\nabla_x \ell_x(Y,D) - \nabla_x \ell_x(Y',D)\| \\ &\le G_{\mathrm{M}}\left[\omega_D + (1-\omega_D)\max\{\varepsilon_{\alpha,\gamma}, t_D\}\right] = G_{\mathrm{M}} v_D.\end{aligned}$$

For $\widehat{\mathcal{C}}_\gamma(D) = \varnothing$, the same conclusion follows from the global bound and $v_D = 1$. Thus, $x \mapsto \mathbb{E}_{y\sim P_{\mathrm{M},\boldsymbol{u}}(\cdot|D)}[\ell_x(y,D)] - g_D(x)$ is $G_{\mathrm{M}} v_D$-Lipschitz on $\mathcal{X}(D)$. Taking the supremum over $\boldsymbol{u} \in \widehat{\mathcal{U}}_\alpha(D;\gamma)$ preserves this bound, giving

$$|(f_D - g_D)(x) - (f_D - g_D)(z)| \le G_{\mathrm{M}} v_D \|x - z\|, \qquad x, z \in \mathcal{X}(D). \tag{59}$$

By Lemma 4 and the curvature argument in the proof of Theorem 6, both $f_D$ and $g_D$ are $\kappa_{\mathrm{M}}$-strongly convex on $\mathcal{X}(D)$. Their respective minimizers are $\bar{x}(D)$ and $x^\star(D)$, with $g_D(x^\star(D)) = 0$. Writing $d_D = \|\bar{x}(D) - x^\star(D)\|$, quadratic growth gives

$$g_D(\bar{x}(D)) \ge \frac{\kappa_{\mathrm{M}}}{2} d_D^2, \qquad f_D(x^\star(D)) - f_D(\bar{x}(D)) \ge \frac{\kappa_{\mathrm{M}}}{2} d_D^2.$$

On the other hand, (59) implies

$$g_D(\bar{x}(D)) + f_D(x^\star(D)) - f_D(\bar{x}(D)) \le G_{\mathrm{M}} v_D d_D.$$

Subtracting either quadratic lower bound from this inequality and completing the square yields

$$\max\{g_D(\bar{x}(D)), f_D(x^\star(D)) - f_D(\bar{x}(D))\} \le G_{\mathrm{M}} v_D d_D - \frac{\kappa_{\mathrm{M}}}{2} d_D^2 \le \frac{G_{\mathrm{M}}^2}{2\kappa_{\mathrm{M}}} v_D^2. \tag{60}$$

Taking expectations and using the definitions of $b_{\alpha,\gamma}$ and $\Xi_{\alpha,\gamma}$ yields

$$b_{\alpha,\gamma} = \mathbb{E}_D[g_D(\bar{x}(D))] \le \frac{G_{\mathrm{M}}^2}{2\kappa_{\mathrm{M}}} \mathbb{E}_D[v_D^2] = \frac{G_{\mathrm{M}}^2}{2\kappa_{\mathrm{M}}} \Xi_{\alpha,\gamma}.$$

This proves the second bound in (47). Moreover, the outcome-wise loss bound on $\mathcal{X}(D)$ established in the proof of Lemma 3 implies $g_D(\bar{x}(D)) \le H_{\mathrm{M}}$. Combining these two bounds proves (47).

Finally, the proof of Theorem 5 establishes global $\Lambda_{\mathrm{M}}$-smoothness of $g_D$. Since $g_D$ is nonnegative and vanishes at $x^\star(D)$, $\nabla g_D(x^\star(D)) = 0$, and hence

$$g_D(x) \le \frac{\Lambda_{\mathrm{M}}}{2} \|x - x^\star(D)\|^2.$$

For every $\pi \in \Pi$, decompose the excess pessimistic loss as

$$e_\pi(D) = g_D(x_\pi(D)) + (f_D - g_D)(x_\pi(D)) - (f_D - g_D)(x^\star(D)) + f_D(x^\star(D)) - f_D(\bar{x}(D)).$$

Combining the preceding smoothness bound with (59) and (60) yields

$$e_\pi(D) \le \frac{\Lambda_{\mathrm{M}}}{2} \|x_\pi(D) - x^\star(D)\|^2 + G_{\mathrm{M}} v_D \|x_\pi(D) - x^\star(D)\| + \frac{G_{\mathrm{M}}^2}{2\kappa_{\mathrm{M}}} v_D^2.$$

Taking expectations and applying Cauchy–Schwarz gives

$$\begin{aligned}\mathbb{E}_D[e_\pi(D)] \le{}& \frac{\Lambda_{\mathrm{M}}}{2} \mathbb{E}_D\left[\|x_\pi(D) - x^\star(D)\|^2\right] \\ &+ G_{\mathrm{M}} \sqrt{\Xi_{\alpha,\gamma} \mathbb{E}_D\left[\|x_\pi(D) - x^\star(D)\|^2\right]} + \frac{G_{\mathrm{M}}^2}{2\kappa_{\mathrm{M}}} \Xi_{\alpha,\gamma}.\end{aligned}$$

Since $\mathcal{E}_\Pi^{\mathrm{pess}} \le \mathbb{E}_D[e_\pi(D)]$ for every $\pi \in \Pi$, applying this bound along a sequence of policies whose mean squared score errors converge to $\mathcal{E}_\Pi$ and passing to the limit proves (48). □

## E Worst-Case Lower Bound

We give the full statement and proof of Theorem 3.

**Theorem 3** (Worst-case Lower Bound for LatentCP-DPO, full statement)**.** *Fix $\alpha \in (0,1)$, $B, \beta > 0$, and $\gamma \in (0, \alpha/2)$ independently of the calibration sample. For every integer $d_\Pi \ge 1$, and all sample sizes satisfying*

$$n_{\mathrm{pil}} \ge 2d_\Pi + 1, \qquad n_{\mathrm{cal}} \ge \frac{1}{\gamma} - 1, \qquad n_{\mathrm{opt}} \ge 8 \max\left\{d_\Pi, \frac{1}{\gamma}\right\},$$

*one can construct a BT instance and a log-linear policy class* $\Pi$ *with a fixed offset and* $d_\Pi + 1$ *free parameters such that Assumptions 1 and 2 hold, and*

$$\pi^\star \in \Pi, \qquad \mathcal{E}_\Pi = 0, \qquad C_{\mathrm{comp}}(\pi^\star; \Pi) \le 1.$$

*Moreover, the localized-complexity condition (26) holds with complexity parameter* $d_\Pi$ *and* $C_{\mathrm{loc}} = \sqrt{3}$*. For this instance, there exists a pilot-sample event* $\mathcal{H}_{\mathrm{pil}}$ *of positive probability such that, conditional on* $\mathcal{H}_{\mathrm{pil}}$*, with probability at least* $1/16$ *over the calibration and optimization samples, every exact LatentCP-DPO empirical optimizer satisfies*

$$\mathrm{SubOpt}_\beta(\widehat{\pi}_{\mathrm{LCP}}, \pi^\star) \ge c_{B,\beta} \left[ \frac{d_\Pi}{n_{\mathrm{opt}}} + \left(1 - \frac{2\gamma}{\alpha}\right)^2 \left(\gamma + \frac{1}{n_{\mathrm{cal}}}\right) \right], \tag{18}$$

*where* $c_{B,\beta} > 0$ *depends only on* $B$ *and* $\beta$*. One valid choice is*

$$c_{B,\beta} = \min\left\{ \frac{v_{B,\beta} a_B^2}{256\beta}, \frac{v_{B,\beta} \min\{B^2, 4\}}{16\beta} \right\}, \tag{61}$$

*with*

$$v_{B,\beta} = \sigma(B/\beta)\sigma(-B/\beta), \qquad a_B = \frac{\sigma(-B/4) - \sigma(-B/2)}{2}.$$

**Proof of Theorem 3.** We organize the proof into six steps, constructing a single instance in which calibration-induced bias and optimization-sample fluctuations contribute additively to policy suboptimality.

**Step 1: Construction of the instance.** Set

$$j = \lfloor (n_{\mathrm{cal}} + 1)\gamma \rfloor, \qquad w = \min\left\{ \frac{1}{4}, \frac{j}{2n_{\mathrm{cal}}} \right\}, \qquad h = \frac{B}{16}.$$

The sample-size assumption gives $j \ge 1$. Choose $\ell \in \{B/4, B/2\}$ such that

$$|\varepsilon_{\alpha,\gamma} - \sigma(-\ell)| \ge a_B.$$

Such a choice exists because

$$2a_B = \sigma(-B/4) - \sigma(-B/2) \le |\varepsilon_{\alpha,\gamma} - \sigma(-B/4)| + |\varepsilon_{\alpha,\gamma} - \sigma(-B/2)|.$$

Since $\sup_{z \ge 0} \sigma'(z) \le 1/4$, we have

$$a_B \le \frac{1}{2} \sup_{z \ge 0} \sigma'(z) \cdot (-B/4 + B/2) = \frac{B}{32} = \frac{h}{2}.$$

Let the prompt space be

$$\mathcal{S} = \{s_c, s_0\} \cup \{s_{i,+}, s_{i,-} : 1 \le i \le d_\Pi\}.$$

Every prompt has two responses, 0 and 1, and the comparison context is the fixed pair $D = (s, 0, 1)$. Take $\pi_{\mathrm{ref}}(0 \mid s) = \pi_{\mathrm{ref}}(1 \mid s) = 1/2$. For

$$\vartheta = (t, \theta_1, \ldots, \theta_{d_\Pi}) \in [0, B] \times [-h, h]^{d_\Pi},$$

define the prompt probabilities and comparison scores by

| $s$ | $\rho(s)$ | $x_\vartheta(s)$ | $x^\star(s)$ |
|---|---|---|---|
| $s_c$ | $w$ | $t$ | $0$ |
| $s_{i,+}$ | $1/(8d_\Pi)$ | $\ell + \theta_i$ | $\ell$ |
| $s_{i,-}$ | $1/(8d_\Pi)$ | $-\ell + \theta_i$ | $-\ell$ |
| $s_0$ | $3/4 - w$ | $0$ | $0$ |

where $w = \rho(s_c)$ is defined above and $i = 1, \ldots, d_\Pi$. Under the uniform reference policy, define

$$\pi_\vartheta(1 \mid s) = \sigma\left(\frac{x_\vartheta(s)}{\beta}\right), \qquad r^\star(s, a) = \frac{2a - 1}{2} x^\star(s).$$

Then the policy-induced comparison score satisfies

$$\beta \log \frac{\pi_\vartheta(1 \mid s)/\pi_{\mathrm{ref}}(1 \mid s)}{\pi_\vartheta(0 \mid s)/\pi_{\mathrm{ref}}(0 \mid s)} = x_\vartheta(s).$$

The comparison score is affine in $\vartheta = (t, \theta_1, \ldots, \theta_{d_\Pi})$, with fixed offsets $\pm\ell$ on the paired prompts. Representing these offsets by an additional feature whose coefficient is fixed at one yields a log-linear policy class with $d_\Pi + 1$ free parameters.

The true BT labels satisfy

$$\mathbb{P}\{y = 1 \mid s\} = \frac{e^{r^\star(s,1)}}{e^{r^\star(s,0)} + e^{r^\star(s,1)}} = \sigma\left(r^\star(s,1) - r^\star(s,0)\right) = \sigma(x^\star(s)).$$

Setting all free parameters in $\vartheta$ to zero gives $x_{\mathbf{0}}(s) = x^\star(s)$ for every prompt. Under the uniform reference policy, (2) therefore yields $\pi^\star = \pi_{\mathbf{0}}$. Draw all pilot and calibration preference units independently from this distribution, and draw the optimization contexts independently from its context marginal.

Enumerate the pilot observations as $(s_i^{\mathrm{pil}}, y_i^{\mathrm{pil}})$, $i = 1, \ldots, n_{\mathrm{pil}}$. Define the event

$$\mathcal{H}_{\mathrm{pil}} = \bigcap_{i=1}^{d_\Pi} \left\{(s_{2i-1}^{\mathrm{pil}}, y_{2i-1}^{\mathrm{pil}}) = (s_{i,+}, 1), (s_{2i}^{\mathrm{pil}}, y_{2i}^{\mathrm{pil}}) = (s_{i,-}, 0)\right\} \cap \bigcap_{r=2d_\Pi+1}^{n_{\mathrm{pil}}} \left\{(s_r^{\mathrm{pil}}, y_r^{\mathrm{pil}}) = (s_c, 1)\right\}.$$

By independence and the specified BT label probabilities,

$$\mathbb{P}(\mathcal{H}_{\mathrm{pil}}) = \left(\frac{\sigma(\ell)}{8d_\Pi}\right)^{2d_\Pi} \left(\frac{w}{2}\right)^{n_{\mathrm{pil}} - 2d_\Pi} > 0.$$

On this event, the empirical pilot loss is

$$\widehat{L}_{\mathrm{pil}}(\vartheta) = \frac{n_{\mathrm{pil}} - 2d_\Pi}{n_{\mathrm{pil}}} \log(1 + e^{-t}) + \frac{1}{n_{\mathrm{pil}}} \sum_{i=1}^{d_\Pi} \left[\log(1 + e^{-\ell - \theta_i}) + \log(1 + e^{-\ell + \theta_i})\right].$$

Since $n_{\mathrm{pil}} \geq 2d_\Pi + 1$, the coefficient of $\log(1 + e^{-t})$ is positive, and this term is strictly decreasing in $t \in [0, B]$. Its unique minimizer is therefore $t = B$. For each $i$, the corresponding two terms form a strictly convex even function of $\theta_i$, whose unique minimizer on $[-h, h]$ is zero. Consequently, $\{(B, 0, \ldots, 0)\} = \operatorname{argmin}_\vartheta \widehat{L}_{\mathrm{pil}}(\vartheta)$.

**Step 2: Verification of the assumptions and complexity bound.** The sampling assumption holds by construction. Moreover, $|x_\vartheta(s)| \leq B$ and $|x^\star(s)| \leq B$ for every prompt and parameter, because $\ell + h \leq 9B/16$. The parameter domain is compact, and the empirical pilot and pessimistic losses are continuous, so their minima are attained. All policies have full response support, and $\pi^\star \in \Pi$ gives $\mathcal{E}_\Pi = 0$.

For any two responses $a, a' \in \{0, 1\}$, the difference of implicit-reward errors is either zero or $\pm(x_\vartheta(s) - x^\star(s))$. Taking expectations over $s \sim \rho$ and conditionally independent responses $a \sim \pi^\star(\cdot \mid s)$ and $a' \sim \pi_\vartheta(\cdot \mid s)$ gives

$$\begin{aligned}
&\mathbb{E}_{\substack{s\sim\rho \\ a\sim\pi^\star(\cdot|s), a'\sim\pi_\vartheta(\cdot|s)}} \left[\left([r_{\pi_\vartheta}(s,a) - r^\star(s,a)] - [r_{\pi_\vartheta}(s,a') - r^\star(s,a')]\right)^2\right] \\
&\qquad \leq \mathbb{E}_{s\sim\rho}\left[|x_\vartheta(s) - x^\star(s)|^2\right] = \mathbb{E}_D\left[|x_{\pi_\vartheta}(D) - x^\star(D)|^2\right],
\end{aligned}$$

where the equality uses the deterministic comparison pair $D = (s, 0, 1)$. Thus, every ratio in the supremum defining $C_{\mathrm{comp}}(\pi^\star; \Pi)$ is at most one. Taking the supremum over $\pi_\vartheta \in \Pi$ therefore yields $C_{\mathrm{comp}}(\pi^\star; \Pi) \leq 1$.

To verify localized complexity, condition on any pilot and calibration realization. At $s_0$, every policy has score zero, so the prediction set is either empty or full, its pointwise pessimistic benchmark is zero, and

$$e_\pi(s_0) = f_{s_0}(x_\pi(s_0)) - f_{s_0}(\bar{x}(s_0)) = f_{s_0}(0) - f_{s_0}(0) = 0.$$

Consequently, every $g \in \mathcal{H}_\zeta$ vanishes at $s_0$ and can be nonzero only at the other $2d_\Pi + 1$ prompts. By the definition of $\mathcal{H}_\zeta$, it also satisfies $0 \leq g \leq 1$ and $\mathbb{E}_\rho[g] \leq \zeta$.

Let

$$Z_s = \frac{1}{n_{\mathrm{opt}}} \sum_{r=1}^{n_{\mathrm{opt}}} \epsilon_r \mathbb{I}\{s_r^{\mathrm{opt}} = s\}, \qquad \mathcal{S}_+ = \mathcal{S} \setminus \{s_0\},$$

where the $\epsilon_r$ are independent Rademacher signs. Weighted Cauchy–Schwarz gives

$$\sup_{g \in \mathcal{H}_\zeta} \left| \sum_{s \in \mathcal{S}_+} g(s) Z_s \right| \le \sup_{g \in \mathcal{H}_\zeta} \left| \sum_{s \in \mathcal{S}_+} \left[\sqrt{\rho(s)} g(s)\right] \left[\frac{Z_s}{\sqrt{\rho(s)}}\right] \right| \le \sqrt{\zeta} \left( \sum_{s \in \mathcal{S}_+} \frac{Z_s^2}{\rho(s)} \right)^{1/2},$$

using $\sum_s \rho(s) g(s)^2 \le \mathbb{E}_\rho[g] \le \zeta$. Since $\mathbb{E}[Z_s^2] = \rho(s)/n_{\mathrm{opt}}$, Jensen's inequality yields

$$\mathfrak{R}_{n_{\mathrm{opt}}}(\zeta) \le \sqrt{\frac{(2d_\Pi + 1)\zeta}{n_{\mathrm{opt}}}} \le \sqrt{3} \sqrt{\frac{d_\Pi \zeta}{n_{\mathrm{opt}}}}.$$

This proves (26) with $C_{\mathrm{loc}} = \sqrt{3}$ and complexity parameter $d_\Pi$.

**Step 3: Policy-loss decomposition.** Define

$$L_\beta(x, z) = \beta D_{\mathrm{KL}} \left( \mathrm{Bern}(\sigma(x/\beta)) \| \mathrm{Bern}(\sigma(z/\beta)) \right).$$

For $F(u) = \log(1 + e^u)$,

$$\frac{L_\beta(x, z)}{\beta} = F(z/\beta) - F(x/\beta) - F'(x/\beta) \frac{z - x}{\beta}.$$

Because $F''(u) \ge v_{B,\beta}$ on $[-B/\beta, B/\beta]$, it follows that

$$L_\beta(x, z) \ge \frac{v_{B,\beta}}{2\beta} (x - z)^2, \qquad x, z \in [-B, B].$$

The Gibbs-policy identity used in the proof of Theorem 6 gives

$$\begin{aligned} \mathrm{SubOpt}_\beta(\pi_\vartheta, \pi^\star) =& \beta \mathbb{E}_{s \sim \rho} \left[ D_{\mathrm{KL}}\big(\pi_\vartheta(\cdot \mid s) \| \pi^\star(\cdot \mid s)\big) \right] \\ =& \mathbb{E}_{s \sim \rho} \left[ L_\beta(x_\vartheta(s), x^\star(s)) \right] \\ =& \sum_{s \in \mathcal{S}} \rho(s) L_\beta(x_\vartheta(s), x^\star(s)) \\ =& w L_\beta(t, 0) + \frac{1}{8 d_\Pi} \sum_{i=1}^{d_\Pi} \left[ L_\beta(\ell + \theta_i, \ell) + L_\beta(-\ell + \theta_i, -\ell) \right] \\ \ge& w L_\beta(t, 0) + \frac{v_{B,\beta}}{8\beta d_\Pi} \sum_{i=1}^{d_\Pi} \theta_i^2 \end{aligned} \tag{62}$$

For later use, write

$$c_\gamma = \min \left\{ B, \log \frac{1 - \varepsilon_{\alpha,\gamma}}{\varepsilon_{\alpha,\gamma}} \right\}, \qquad A_\gamma = L_\beta(c_\gamma, 0).$$

In particular,

$$A_\gamma = \frac{c_\gamma}{2} \tanh\left(\frac{c_\gamma}{2\beta}\right) - \beta \log\cosh\left(\frac{c_\gamma}{2\beta}\right) \ge \frac{v_{B,\beta}}{2\beta} c_\gamma^2 \ge \frac{v_{B,\beta} \min\{B^2, 4\}}{2\beta} (1 - 2\varepsilon_{\alpha,\gamma})^2. \tag{63}$$

Here the last inequality uses $\log((1 - \varepsilon_{\alpha,\gamma})/\varepsilon_{\alpha,\gamma}) \ge 2(1 - 2\varepsilon_{\alpha,\gamma})$ and hence $c_\gamma \ge \min\{B, 2\}(1 - 2\varepsilon_{\alpha,\gamma})$.

**Step 4: LatentCP-induced contribution.** For the remainder of the proof, condition on $\mathcal{H}_{\mathrm{pil}}$. The pilot comparison scores are

$$x_{\widehat{\pi}_{\mathrm{pil}}}(s_c) = B, \qquad x_{\widehat{\pi}_{\mathrm{pil}}}(s_{i,\pm}) = \pm\ell, \qquad x_{\widehat{\pi}_{\mathrm{pil}}}(s_0) = 0.$$

By the BT nonconformity-score definition,

$$\widehat{s}(s, y) = \log\left(1 + \exp\left\{-(2y - 1) x_{\widehat{\pi}_{\mathrm{pil}}}(s)\right\}\right).$$

Since $0 < \ell < B$, the largest possible score is $H_B = \log(1 + e^B)$, attained precisely at $(s_c, 0)$, whereas the smallest is $L_B = \log(1 + e^{-B})$, attained precisely at $(s_c, 1)$. All scores at the other prompts lie strictly between $L_B$ and $H_B$.

Let $N_H$ count calibration observations with score $H_B$. Then

$$N_H \sim \text{Binomial}(n_{\text{cal}}, w/2), \qquad k_\gamma = \lceil (n_{\text{cal}} + 1)(1 - \gamma) \rceil = n_{\text{cal}} + 1 - j \le n_{\text{cal}}.$$

On $\mathcal{G}_{\text{cal}} = \{N_H \le j - 1\}$, the calibrated threshold satisfies $L_B \le \widehat{q}^{\text{cal}}_{1-\gamma} < H_B$, so $\widehat{\mathcal{C}}_\gamma(s_c) = \{1\}$. Markov's inequality gives

$$\mathbb{P}(\mathcal{G}_{\text{cal}} \mid \mathcal{H}_{\text{pil}}) = 1 - \mathbb{P}(N_H \ge j \mid \mathcal{H}_{\text{pil}}) \ge 1 - \frac{\mathbb{E}[N_H \mid \mathcal{H}_{\text{pil}}]}{j} \ge 1 - \frac{n_{\text{cal}} w}{2j} \ge \frac{3}{4}.$$

If the optimization sample contains $s_c$, the empirical objective is uniquely minimized with respect to $t = x_\vartheta(s_c)$ at $t = c_\gamma$, by (13). Its contribution in (62) is therefore $wA_\gamma$.

We next lower-bound $w$ to quantify the contribution of $s_c$ and control its occurrence in the optimization sample. Since $(n_{\text{cal}} + 1)\gamma \ge 1$, the inequality $\lfloor r \rfloor \ge r/2$ for $r \ge 1$ gives

$$\frac{j}{2n_{\text{cal}}} \ge \frac{(n_{\text{cal}} + 1)\gamma}{4n_{\text{cal}}} \ge \frac{\gamma}{4}.$$

Together with $1/4 \ge \gamma/4$, this implies $w \ge \gamma/4$. Similarly, $j \ge 1$ and $n_{\text{cal}} \ge 1$ imply

$$\frac{j}{2n_{\text{cal}}} \ge \frac{1}{4n_{\text{cal}}}, \qquad \frac{1}{4} \ge \frac{1}{4n_{\text{cal}}},$$

so $w \ge 1/(4n_{\text{cal}})$. Combining these bounds yields

$$w \ge \frac{\gamma}{4}, \quad w \ge \frac{1}{4n_{\text{cal}}} \qquad \Rightarrow \quad w \ge \frac{1}{8}\left(\gamma + \frac{1}{n_{\text{cal}}}\right). \tag{64}$$

Consequently, the contribution $wA_\gamma$ is at least $A_\gamma(\gamma + 1/n_{\text{cal}})/8$. Moreover, the sample-size condition $n_{\text{opt}} \ge 8/\gamma$ gives $n_{\text{opt}} w \ge 2$. Since $\gamma < 1/2$, it also gives $n_{\text{opt}} > 16$.

**Step 5: Optimization-sample contribution.** Let $N_c$ count optimization observations at $s_c$, and let $N_{i,+}, N_{i,-}$ count those at $s_{i,+}, s_{i,-}$. Write

$$M_i = N_{i,+} + N_{i,-}, \qquad \Delta_i = N_{i,+} - N_{i,-}.$$

First fix any calibration realization. For each paired group, the pilot scores are $\ell$ and $-\ell$, so the two labels' nonconformity scores are exchanged. Applying the same calibration threshold therefore yields either $\{1\}$ at $s_{i,+}$ and $\{0\}$ at $s_{i,-}$, or prediction sets that are both empty or both full. Since $\ell - h > 0$, the empirical objective for coordinate $\theta_i$ can be written as

$$F_i(\theta) = \frac{1}{n_{\text{opt}}}\left[N_{i,+}\psi(\ell + \theta) + N_{i,-}\psi(\ell - \theta)\right],$$

where, on the positive score range under consideration,

$$\psi(z) = \begin{cases} \log(1 + e^{-z}) + \varepsilon_{\alpha,\gamma} z, & \text{for singleton sets,} \\ \log(1 + e^z), & \text{for empty or full sets.} \end{cases}$$

Thus, if $M_i > 0$,

$$F_i'(0) = \frac{\Delta_i \psi'(\ell)}{n_{\text{opt}}}, \qquad 0 < F_i''(\theta) \le \frac{M_i}{4n_{\text{opt}}},$$

where $\psi'(\ell) = \varepsilon_{\alpha,\gamma} - \sigma(-\ell)$ in the singleton case and $\psi'(\ell) = \sigma(\ell)$ otherwise. By the choice of $\ell$ in the former case and $\sigma(\ell) > 1/2 > a_B$ in the latter, we have $|\psi'(\ell)| \ge a_B$ in both cases.

Fix $i$ with $M_i > 0$. If an empirical minimizer satisfies $|\widehat{\theta}_i| < h$, first-order optimality gives $F_i'(\widehat{\theta}_i) = 0$. The curvature upper bound therefore implies

$$\frac{a_B|\Delta_i|}{n_{\text{opt}}} \le |F_i'(0)| = |F_i'(0) - F_i'(\widehat{\theta}_i)| \le \frac{M_i}{4n_{\text{opt}}}|\widehat{\theta}_i|,$$

and hence $|\widehat{\theta}_i| \geq 4a_B|\Delta_i|/M_i$. If the minimizer lies on the boundary of $[-h, h]$, then $|\widehat{\theta}_i| = h$. Combining the two cases yields, for every empirical minimizer,

$$|\widehat{\theta}_i| \geq \min\left\{h, \frac{4a_B|\Delta_i|}{M_i}\right\}.$$

Since $|\Delta_i| \leq M_i$ and $h \geq 2a_B$, this yields

$$\widehat{\theta}_i^2 \geq 4a_B^2 \frac{\Delta_i^2}{M_i^2}, \qquad M_i > 0. \tag{65}$$

We next construct an optimization event on which these errors accumulate. Let $\mathcal{B} = \{1 \leq N_c \leq n_{\mathrm{opt}}/2\}$. Because $N_c \sim \mathrm{Binomial}(n_{\mathrm{opt}}, w)$, $n_{\mathrm{opt}} w \geq 2$, and $w \leq 1/4$,

$$\mathbb{P}(N_c = 0) = (1-w)^{n_{\mathrm{opt}}} \leq e^{-2}, \quad \mathbb{P}(N_c > n_{\mathrm{opt}}/2) \leq \frac{\mathbb{E}[3^{N_c}]}{3^{n_{\mathrm{opt}}/2}} \leq \left(\frac{\sqrt{3}}{2}\right)^{n_{\mathrm{opt}}} \leq \left(\frac{3}{4}\right)^8.$$

Hence $\mathbb{P}(\mathcal{B}) > 3/4$. Fix an integer $k$ with $1 \leq k \leq n_{\mathrm{opt}}/2$ and condition on $N_c = k$. Then each $M_i$ is binomial with mean

$$\mu_k = \frac{n_{\mathrm{opt}} - k}{4d_\Pi(1-w)} \geq \frac{n_{\mathrm{opt}}}{8d_\Pi} \geq 1.$$

Let $\mathcal{I}_M = \{i : M_i > 0\}$ and $g = |\mathcal{I}_M|$. For each $i$, the conditional binomial distribution gives

$$\mathbb{P}(M_i = 0 \mid N_c = k) = \left(1 - \frac{1}{4d_\Pi(1-w)}\right)^{n_{\mathrm{opt}}-k} \leq e^{-\mu_k} \leq e^{-1}.$$

Since $g = \sum_{i=1}^{d_\Pi} \mathbb{I}\{M_i > 0\}$, linearity of expectation yields

$$\mathbb{E}[g \mid N_c = k] = \sum_{i=1}^{d_\Pi} \mathbb{P}(M_i > 0 \mid N_c = k) \geq d_\Pi(1 - e^{-1}) \geq \frac{5d_\Pi}{8}.$$

On the other hand, using $0 \leq g \leq d_\Pi$ and splitting according to whether $g \geq d_\Pi/4$, we obtain

$$\begin{aligned}\mathbb{E}[g \mid N_c = k] &\leq \frac{d_\Pi}{4}\mathbb{P}(g < d_\Pi/4 \mid N_c = k) + d_\Pi \mathbb{P}(g \geq d_\Pi/4 \mid N_c = k) \\ &= \frac{d_\Pi}{4} + \frac{3d_\Pi}{4}\mathbb{P}(g \geq d_\Pi/4 \mid N_c = k).\end{aligned}$$

Combining the preceding two bounds and rearranging gives

$$\mathbb{P}(g \geq d_\Pi/4 \mid N_c = k) \geq \frac{1}{2}.$$

Conditional on $N_c$ and $(M_i)_{i=1}^{d_\Pi}$, the number of observations in each paired group is fixed. Since $\rho(s_{i,+}) = \rho(s_{i,-})$, each observation within group $i$ is independently drawn from $s_{i,+}$ or $s_{i,-}$ with equal probability. These within-group allocations are also independent across groups. Consequently,

$$N_{i,+} \mid N_c, (M_i)_{i=1}^{d_\Pi} \sim \mathrm{Binomial}(M_i, 1/2), \qquad \Delta_i = 2N_{i,+} - M_i. \tag{66}$$

Thus, the $\Delta_i$ are conditionally independent, and each is the sum of $M_i$ independent signs, with $+1$ representing an occurrence of $s_{i,+}$ and $-1$ an occurrence of $s_{i,-}$.

Fix a realization of $N_c$ and $(M_i)_{i=1}^{d_\Pi}$ for which $g \geq d_\Pi/4$, and define

$$Z = \sum_{i \in \mathcal{I}_M} \frac{\Delta_i^2}{M_i^2}.$$

Conditional on these counts, by (66), we have,

$$\mathbb{E}[\Delta_i \mid N_c, (M_i)] = 0, \qquad \mathrm{Var}(\Delta_i \mid N_c, (M_i)) = M_i,$$

and hence $\mathbb{E}[\Delta_i^2 \mid N_c, (M_i)] = M_i$. Expanding the fourth power of the sum of the independent $\pm 1$ contributions gives

$$\mathbb{E}[\Delta_i^4 \mid N_c, (M_i)] = M_i + 6\binom{M_i}{2} = 3M_i^2 - 2M_i \le 3M_i^2,$$

because only terms in which every sign appears an even number of times have nonzero expectation. Consequently, the conditional mean of $Z$ is

$$L = \mathbb{E}[Z \mid N_c, (M_i)] = \sum_{i \in \mathcal{I}_M} \frac{1}{M_i}.$$

Expanding $Z^2$ and using conditional independence across groups, we obtain

$$\begin{aligned}\mathbb{E}[Z^2 \mid N_c, (M_i)] &= \sum_{i \in \mathcal{I}_M} \frac{\mathbb{E}[\Delta_i^4 \mid N_c, (M_i)]}{M_i^4} + 2 \sum_{\substack{i<j \\ i,j \in \mathcal{I}_M}} \frac{1}{M_i M_j} \\ &\le 3 \sum_{i \in \mathcal{I}_M} \frac{1}{M_i^2} + 2 \sum_{\substack{i<j \\ i,j \in \mathcal{I}_M}} \frac{1}{M_i M_j} \\ &\le 3 \left( \sum_{i \in \mathcal{I}_M} \frac{1}{M_i} \right)^2 = 3L^2.\end{aligned}$$

Here the equality uses $\mathbb{E}[\Delta_i^2 \Delta_j^2 \mid N_c, (M_i)] = M_i M_j$ for $i \neq j$, and the first inequality uses the fourth-moment bound above. Moreover, Cauchy–Schwarz gives

$$g^2 = \left( \sum_{i \in \mathcal{I}_M} \sqrt{M_i} \frac{1}{\sqrt{M_i}} \right)^2 \le \left( \sum_{i \in \mathcal{I}_M} M_i \right) L.$$

Since $\sum_{i \in \mathcal{I}_M} M_i \le n_{\mathrm{opt}}$ and $g \ge d_\Pi/4$, we conclude that

$$L \ge \frac{g^2}{\sum_{i \in \mathcal{I}_M} M_i} \ge \frac{g^2}{n_{\mathrm{opt}}} \ge \frac{d_\Pi^2}{16 n_{\mathrm{opt}}}.$$

Paley–Zygmund inequality therefore yields

$$\mathbb{P}\{Z \ge L/8 \mid N_c, (M_i)\} \ge (1 - 1/8)^2 \frac{L^2}{3L^2} = \frac{49}{192}.$$

Consequently, there is an event $\mathcal{G}_{\mathrm{opt}} = \mathcal{B} \cap \{g \ge d_\Pi/4\} \cap \{Z \ge L/8\}$, depending only on the optimization contexts, with

$$\mathbb{P}(\mathcal{G}_{\mathrm{opt}}) \ge \frac{3}{4} \cdot \frac{1}{2} \cdot \frac{49}{192} = \frac{49}{512},$$

on which $N_c \ge 1$ and $Z \ge d_\Pi^2/(128 n_{\mathrm{opt}})$. Combining (65) with (62), the statistical contribution on this event is at least

$$\frac{v_{B,\beta}}{8\beta d_\Pi} \sum_{i=1}^{d_\Pi} \widehat{\theta}_i^2 \ge \frac{v_{B,\beta}}{8\beta d_\Pi} \sum_{i \in \mathcal{I}_M} \widehat{\theta}_i^2 \ge \frac{v_{B,\beta} a_B^2}{2\beta d_\Pi} \sum_{i \in \mathcal{I}_M} \frac{\Delta_i^2}{M_i^2} = \frac{v_{B,\beta} a_B^2}{2\beta d_\Pi} Z \ge \frac{v_{B,\beta} a_B^2}{256\beta} \frac{d_\Pi}{n_{\mathrm{opt}}}.$$

This statement holds for every calibration realization and every exact empirical optimizer.

**Step 6: Combining the contributions.** Conditional on $\mathcal{H}_{\mathrm{pil}}$, the events $\mathcal{G}_{\mathrm{cal}}$ and $\mathcal{G}_{\mathrm{opt}}$ are independent. Their intersection has conditional probability at least

$$\frac{3}{4} \cdot \frac{49}{512} = \frac{147}{2048} > \frac{1}{16}.$$

On this intersection, every exact empirical optimizer has $\widehat{t} = c_\gamma$, and the preceding bounds give

$$\begin{aligned}\mathrm{SubOpt}_\beta(\widehat{\pi}_{\mathrm{LCP}}, \pi^\star) &\ge \frac{v_{B,\beta} a_B^2}{256\beta} \frac{d_\Pi}{n_{\mathrm{opt}}} + w A_\gamma \\ &\ge \frac{v_{B,\beta} a_B^2}{256\beta} \frac{d_\Pi}{n_{\mathrm{opt}}} + \frac{v_{B,\beta} \min\{B^2, 4\}}{16\beta} (1 - 2\varepsilon_{\alpha,\gamma})^2 \left( \gamma + \frac{1}{n_{\mathrm{cal}}} \right),\end{aligned}$$

where the last inequality uses (63). Taking the minimum of the two displayed coefficients proves the theorem with the stated $c_{B,\beta}$. □

# F Additional Experimental Results

## F.1 Full Results for the Main Experiment

Table 1 reports each method's average split-conformal prediction-set size at all 50 nominal coverage levels $\tau \in [0.75, 0.95]$. Figure 2 shows five representative levels for legibility, whereas the reported average set-size reductions use all 50. The gains are not driven by a narrow coverage range: LatentCP-DPO yields the smallest set at 49 of the 50 levels, outperforms each individual baseline at no fewer than 49 levels, and never ranks below second.

All evaluated methods share an identical experimental setup: the same base model, corpus, optimizer budget, $1,500$-context calibration split, and $3,000$-context test split. The pilot policy and the calibrated threshold $\widehat{q}$ construct $\widehat{C}$ over all $8,000$ optimization contexts; each final policy is then trained on a $4,000$-context subsample of that pool (i.e., $n_{\text{opt}} = 4,000$). Within a seed, this subsample and the LoRA initialization are common to all methods. We denote CFA as CFA-style$^{\dagger}$ because the original method constructs conformal sets over freshly generated answers. Since our fixed $K = 6$ ranking corpus involves no generation, we adapt CFA by applying its set-wise aggregation directly over the six provided candidates.

## F.2 Cross-Architecture Evaluation

To assess whether the efficiency gains generalize across base models, we evaluate LatentCP-DPO on Llama-3.1-8B-Instruct and Llama-2-7B-Chat. We use the same data, seed assignments, and optimization budgets as in the Qwen3-8B experiment, including a PL pilot trained on $3,000$ contexts and final policies trained on $4,000$ contexts. We transfer $\gamma$ unchanged from the Qwen3-8B validation selection, and no architecture-specific tuning is performed. As shown in Figure 3, LatentCP-DPO consistently yields smaller conformal prediction sets than Pure DPO at a common requested coverage level across all three models. The gains thus persist across the tested architectures without architecture-specific hyperparameter search.

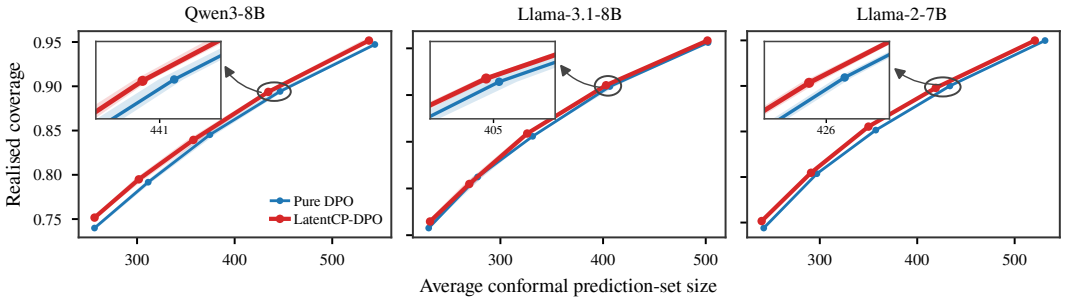


Figure 3: Prediction-set efficiency across base models. Realized coverage versus average conformal ranking prediction-set size for LatentCP-DPO and Pure DPO on three base models under the same experimental protocol. The Llama runs retain the Qwen3-8B values of $\gamma$. Curves show means over three matched policy seeds; shaded bands show their min–max ranges.

Table 1: Full 50-level prediction-set-size results for the fixed Qwen3-8B setting ($K = 6$). Each row corresponds to a requested coverage level $\tau$; entries give the average split-conformal prediction-set size out of 720 possible rankings, averaged over matched policy seeds, with the smallest entry in each row shown in bold. LatentCP-DPO yields the smallest set at **49 of the 50 levels**. The final column reports its reduction relative to Pure DPO, with a bar proportional to that reduction.

| | Average prediction-set size | | | | | | |
|---|---|---|---|---|---|---|---|
| $\tau$ | **Ours** | Pure DPO | CFA$^{\dagger}$ | ROPO | DPOP | SimPO | $\Delta$ vs. Pure DPO |
| 0.7500 | **256.4** | **256.4** | 262.0 | 285.8 | 319.1 | 371.2 | −0.02% |
| 0.7541 | **258.9** | 260.1 | 264.7 | 290.3 | 323.8 | 375.8 | 0.47% |
| 0.7582 | **261.7** | 265.7 | 269.8 | 295.1 | 330.4 | 379.8 | 1.54% |
| 0.7622 | **266.1** | 269.7 | 272.8 | 298.2 | 333.1 | 383.2 | 1.36% |
| 0.7663 | **270.3** | 273.5 | 276.6 | 303.6 | 337.6 | 387.2 | 1.18% |
| 0.7704 | **274.5** | 277.8 | 280.3 | 307.5 | 341.8 | 390.1 | 1.17% |
| 0.7745 | **276.8** | 284.5 | 286.0 | 311.0 | 345.6 | 393.9 | 2.69% |
| 0.7786 | **281.4** | 288.5 | 290.4 | 316.5 | 348.9 | 400.5 | 2.47% |
| 0.7827 | **285.2** | 292.1 | 295.7 | 321.0 | 353.9 | 405.3 | 2.38% |
| 0.7867 | **288.7** | 297.0 | 300.3 | 326.6 | 357.1 | 409.8 | 2.78% |
| 0.7908 | **293.1** | 301.8 | 304.7 | 332.4 | 361.0 | 413.1 | 2.89% |
| 0.7949 | **296.1** | 307.3 | 309.7 | 337.8 | 367.3 | 417.0 | 3.65% |
| 0.7990 | **301.0** | 311.0 | 312.6 | 342.0 | 372.0 | 422.1 | 3.22% |
| 0.8031 | **305.4** | 315.6 | 316.4 | 344.7 | 375.8 | 425.8 | 3.22% |
| 0.8071 | **310.0** | 318.8 | 321.2 | 348.2 | 380.3 | 430.6 | 2.79% |
| 0.8112 | **315.2** | 324.9 | 326.2 | 352.6 | 383.6 | 435.0 | 3.00% |
| 0.8153 | **320.6** | 331.5 | 329.9 | 355.3 | 390.0 | 439.2 | 3.30% |
| 0.8194 | **323.9** | 335.4 | 333.8 | 360.8 | 396.3 | 444.4 | 3.45% |
| 0.8235 | **330.0** | 339.4 | 339.3 | 366.8 | 401.2 | 449.8 | 2.76% |
| 0.8276 | **334.0** | 343.8 | 342.5 | 371.6 | 405.5 | 455.8 | 2.86% |
| 0.8316 | **339.7** | 347.5 | 345.6 | 375.4 | 411.8 | 459.5 | 2.24% |
| 0.8357 | **344.2** | 355.4 | 350.8 | 381.3 | 417.2 | 463.4 | 3.14% |
| 0.8398 | **348.4** | 359.4 | 356.4 | 386.1 | 423.1 | 467.3 | 3.05% |
| 0.8439 | **352.5** | 366.0 | 363.8 | 389.7 | 428.9 | 470.5 | 3.67% |
| 0.8480 | **356.7** | 372.1 | 368.3 | 394.4 | 435.5 | 475.0 | 4.16% |
| 0.8520 | **361.5** | 378.4 | 372.6 | 399.9 | 442.9 | 480.7 | 4.46% |
| 0.8561 | **368.0** | 383.6 | 380.0 | 411.1 | 448.1 | 487.8 | 4.07% |
| 0.8602 | **370.6** | 388.3 | 386.7 | 414.5 | 453.2 | 494.6 | 4.57% |
| 0.8643 | **377.5** | 394.6 | 392.9 | 420.4 | 460.8 | 498.4 | 4.35% |
| 0.8684 | **383.6** | 399.2 | 398.4 | 424.9 | 467.2 | 503.2 | 3.91% |
| 0.8724 | **389.1** | 406.6 | 404.2 | 433.1 | 473.8 | 508.7 | 4.31% |
| 0.8765 | **395.2** | 412.5 | 411.1 | 439.1 | 478.5 | 516.6 | 4.19% |
| 0.8806 | **402.3** | 418.3 | 417.3 | 448.4 | 483.8 | 524.1 | 3.82% |
| 0.8847 | **409.2** | 423.0 | 423.8 | 457.0 | 490.2 | 528.3 | 3.28% |
| 0.8888 | **417.7** | 430.2 | 430.0 | 467.7 | 496.7 | 533.9 | 2.89% |
| 0.8929 | **425.3** | 433.6 | 439.2 | 476.3 | 503.6 | 539.9 | 1.90% |
| 0.8969 | **432.7** | 441.7 | 445.0 | 487.2 | 508.2 | 545.5 | 2.03% |
| 0.9010 | **437.1** | 448.4 | 450.4 | 496.6 | 514.2 | 550.9 | 2.53% |
| 0.9051 | **443.1** | 455.1 | 454.6 | 504.0 | 520.5 | 558.7 | 2.64% |
| 0.9092 | **454.2** | 461.5 | 463.1 | 509.4 | 528.7 | 564.3 | 1.56% |
| 0.9133 | **462.0** | 470.1 | 469.5 | 518.7 | 536.8 | 571.1 | 1.73% |
| 0.9173 | **469.0** | 478.7 | 477.7 | 523.3 | 545.7 | 579.6 | 2.02% |
| 0.9214 | **475.0** | 490.8 | 484.6 | 529.4 | 558.5 | 586.0 | 3.22% |
| 0.9255 | **482.9** | 497.3 | 492.9 | 533.7 | 564.1 | 590.9 | 2.91% |
| 0.9296 | **491.2** | 503.1 | 502.3 | 539.3 | 569.3 | 598.1 | 2.37% |
| 0.9337 | **498.8** | 511.4 | 513.6 | 542.5 | 575.5 | 606.9 | 2.47% |
| 0.9378 | **505.5** | 521.6 | 521.6 | 549.5 | 582.2 | 614.4 | 3.08% |
| 0.9418 | **516.2** | 529.8 | 528.4 | 556.7 | 590.4 | 623.1 | 2.56% |
| 0.9459 | **527.2** | 538.2 | 539.8 | 563.2 | 597.4 | 627.0 | 2.04% |
| 0.9500 | **537.6** | 543.7 | 544.0 | 572.4 | 604.1 | 636.4 | 1.11% |
| **Mean** | **372.46** | 383.10 | 383.27 | 414.26 | 446.70 | 488.69 | **2.78%** |

Table 2: Set-size matrix indexed by realized rather than requested coverage (locked Qwen3-8B endpoint, $K = 6$). Robustness companion to Table 1. Each arm is interpolated onto a common grid of 50 realized coverages spanning $[0.7517, 0.9382]$, the range all six arms attain, and entries are the average prediction-set size out of 720 at that realized coverage. The smallest entry in each row is in bold; LatentCP-DPO is smallest at **50 of the 50 points**. The last column is our reduction against Pure DPO.

| | Average prediction-set size | | | | | | |
|---|---|---|---|---|---|---|---|
| Realized cov. | **Ours** | Pure DPO | CFA$^\dagger$ | ROPO | DPOP | SimPO | $\Delta$ vs. Pure DPO |
| 0.7517 | **256.4** | 267.7 | 267.7 | 300.2 | 333.7 | 390.5 | 4.20% |
| 0.7555 | **259.6** | 271.7 | 271.4 | 303.9 | 339.7 | 394.0 | 4.44% |
| 0.7593 | **264.0** | 275.1 | 275.1 | 308.1 | 345.1 | 397.0 | 4.05% |
| 0.7631 | **267.9** | 279.1 | 278.7 | 312.3 | 350.0 | 400.1 | 4.00% |
| 0.7669 | **271.5** | 283.5 | 282.3 | 317.1 | 355.5 | 403.3 | 4.21% |
| 0.7707 | **275.2** | 287.9 | 285.8 | 320.9 | 359.7 | 406.9 | 4.41% |
| 0.7745 | **278.7** | 292.7 | 289.7 | 325.2 | 364.3 | 410.8 | 4.77% |
| 0.7783 | **282.5** | 296.0 | 294.0 | 330.0 | 369.1 | 414.9 | 4.58% |
| 0.7821 | **287.0** | 300.3 | 298.0 | 334.7 | 373.7 | 418.8 | 4.43% |
| 0.7859 | **292.5** | 305.1 | 302.3 | 339.8 | 378.7 | 422.3 | 4.12% |
| 0.7897 | **297.0** | 309.5 | 306.6 | 344.4 | 383.5 | 427.2 | 4.05% |
| 0.7935 | **300.4** | 313.8 | 310.3 | 348.8 | 389.4 | 431.9 | 4.25% |
| 0.7974 | **304.4** | 317.7 | 316.9 | 353.7 | 394.6 | 436.1 | 4.20% |
| 0.8012 | **308.7** | 322.3 | 320.6 | 357.7 | 399.3 | 439.9 | 4.22% |
| 0.8050 | **313.6** | 326.7 | 324.6 | 362.4 | 404.3 | 445.0 | 4.01% |
| 0.8088 | **318.6** | 330.4 | 330.2 | 367.5 | 409.8 | 448.9 | 3.59% |
| 0.8126 | **323.1** | 334.1 | 333.8 | 372.3 | 415.0 | 453.2 | 3.30% |
| 0.8164 | **327.2** | 339.0 | 339.6 | 377.3 | 420.8 | 457.9 | 3.47% |
| 0.8202 | **332.3** | 343.3 | 344.1 | 382.0 | 426.3 | 462.6 | 3.21% |
| 0.8240 | **336.7** | 347.2 | 348.6 | 386.8 | 431.6 | 468.5 | 3.03% |
| 0.8278 | **341.0** | 351.9 | 353.7 | 392.4 | 437.0 | 472.7 | 3.10% |
| 0.8316 | **347.0** | 356.5 | 358.2 | 397.5 | 442.0 | 476.9 | 2.66% |
| 0.8354 | **352.1** | 360.9 | 362.2 | 402.9 | 448.4 | 481.8 | 2.45% |
| 0.8392 | **357.6** | 365.3 | 366.6 | 408.7 | 453.5 | 485.9 | 2.12% |
| 0.8430 | **363.1** | 370.6 | 371.9 | 414.6 | 458.4 | 491.3 | 2.04% |
| 0.8468 | **368.9** | 376.3 | 376.9 | 419.3 | 463.3 | 497.9 | 1.97% |
| 0.8507 | **374.2** | 381.7 | 382.0 | 424.1 | 468.3 | 503.0 | 1.96% |
| 0.8545 | **379.1** | 387.0 | 387.3 | 428.8 | 473.0 | 507.1 | 2.06% |
| 0.8583 | **382.5** | 391.9 | 391.6 | 433.5 | 477.4 | 511.7 | 2.42% |
| 0.8621 | **386.9** | 397.0 | 396.8 | 439.2 | 482.6 | 516.5 | 2.56% |
| 0.8659 | **392.7** | 403.0 | 402.9 | 443.8 | 487.8 | 521.9 | 2.56% |
| 0.8697 | **399.2** | 408.9 | 408.3 | 448.5 | 493.3 | 527.2 | 2.37% |
| 0.8735 | **406.0** | 414.3 | 413.6 | 453.5 | 499.3 | 532.2 | 2.00% |
| 0.8773 | **411.9** | 419.8 | 419.4 | 458.5 | 505.1 | 538.5 | 1.87% |
| 0.8811 | **417.0** | 426.0 | 425.8 | 463.6 | 510.6 | 544.3 | 2.11% |
| 0.8849 | **422.0** | 432.2 | 431.9 | 469.1 | 517.1 | 549.0 | 2.36% |
| 0.8887 | **427.4** | 438.0 | 438.8 | 475.7 | 522.9 | 556.2 | 2.44% |
| 0.8925 | **433.3** | 444.1 | 445.0 | 481.9 | 527.7 | 563.3 | 2.43% |
| 0.8963 | **438.9** | 451.2 | 452.1 | 487.8 | 532.2 | 569.2 | 2.73% |
| 0.9001 | **444.9** | 458.1 | 458.4 | 492.2 | 536.6 | 574.7 | 2.88% |
| 0.9040 | **451.5** | 464.8 | 463.9 | 496.7 | 542.0 | 580.0 | 2.86% |
| 0.9078 | **457.7** | 471.8 | 470.1 | 503.4 | 547.5 | 585.4 | 3.00% |
| 0.9116 | **463.9** | 478.4 | 476.8 | 510.1 | 552.7 | 590.3 | 3.02% |
| 0.9154 | **471.2** | 484.1 | 484.8 | 515.9 | 557.9 | 597.1 | 2.68% |
| 0.9192 | **477.8** | 489.9 | 490.6 | 523.9 | 562.6 | 602.7 | 2.46% |
| 0.9230 | **484.1** | 496.0 | 498.0 | 530.6 | 567.7 | 608.3 | 2.41% |
| 0.9268 | **490.0** | 503.1 | 505.2 | 538.1 | 573.8 | 614.9 | 2.61% |
| 0.9306 | **497.5** | 512.0 | 511.2 | 544.8 | 579.4 | 622.6 | 2.83% |
| 0.9344 | **504.2** | 520.1 | 521.5 | 552.9 | 585.8 | 630.1 | 3.06% |
| 0.9382 | **511.2** | 526.8 | 528.0 | 560.9 | 593.0 | 636.4 | 2.96% |
| **Mean** | **371.04** | 382.50 | 382.27 | 419.15 | 460.85 | 500.38 | **3.00%** |