CO-OIPO: COVERAGE-OPTIMIZED ONLINE IPO FOR PROCESS REWARDS UNDER INTRANSITIVE PREFERENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Process reward models (PRMs) provide signals for evaluating and guiding intermediate steps in multi-step reasoning. Existing PRMs typically adopt a Bradley–Terry (BT) model trained on preference pairs, with the training objective of distinguishing the better from the worse within a pair of samples. However, in multi-hop knowledge graph question answering, multiple paths may correspond to the correct answer, leaving some intermediate steps without a clear distinction between better and worse. Furthermore, multiple preference pairs may form long-chain preference relations, thereby causing a sample to appear in multiple positions within the preference relations, which gives rise to contradictions. To address the above problem, we propose formulating preference learning as a game over reasoning candidates, with a regularized Nash equilibrium as the learning objective. This formulation accommodates a broader class of pairwise preference relations, allowing two samples without clear preference to receive similar exploration probabilities. Furthermore, to make more effective use of a limited comparison budget, we propose Coverage-Optimized Online Identity Preference Optimization (CO-OIPO), which optimizes pair allocation within a sampled candidate pool. We prove that, with sufficient sampling coverage, an expressive enough policy model, and full optimization in each round, CO-OIPO converges at a fixed point corresponding to a regularized Nash equilibrium of the preference game. Experiments show that CO-OIPO achieves the highest performance on most metrics; the controlled experiments further demonstrate more effective utilization of samples without clear preference, and correctness of theoretical analysis.

1 INTRODUCTION

Multi-hop knowledge graph question answering requires models to integrate evidence across multiple reasoning steps. Process reward models (PRMs) can guide such stepwise reasoning, including implicit PRMs that derive process rewards from trajectory-level supervision (Yuan et al., 2025; Wang et al., 2026). In complex question answering, different reasoning paths may all lead to the correct answer, and the quality of intermediate steps is also influenced by multiple factors such as graph relevance and the logical coherence of the chain of thought. Consequently, multiple intermediate steps may exhibit no clear superiority or preference.

Several implicit PRM formulations adopt Direct Preference Optimization (DPO). For example, ImplicitPRM includes DPO as one of several training objectives for learning policy–reference log-ratio process scores (Yuan et al., 2025; Rafailov et al., 2023); DPRM uses preference pairs to learn implicit process rewards for graph reasoning in multi-hop question answering (Wang et al., 2026); and iStar learns implicit step rewards from trajectory preferences using a multi-turn DPO objective (Liu et al., 2026b). Standard DPO training relies on chosen–rejected pairs, where one reasoning trajectory is preferred over another. In multi-hop question answering, such a preference can be difficult to determine for intermediate reasoning trajectories. A binary chosen–rejected pair may therefore fail to reflect their actual relative merits. Beyond this difficulty in constructing preference pairs, DPO also inherits a structural restriction from the Bradley–Terry (BT) model. Each chosen–rejected pair encourages DPO to assign a higher scalar preference score to the chosen trajectory than to the rejected one. When the underlying preferences are cyclic, different training pairs can impose mutually

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

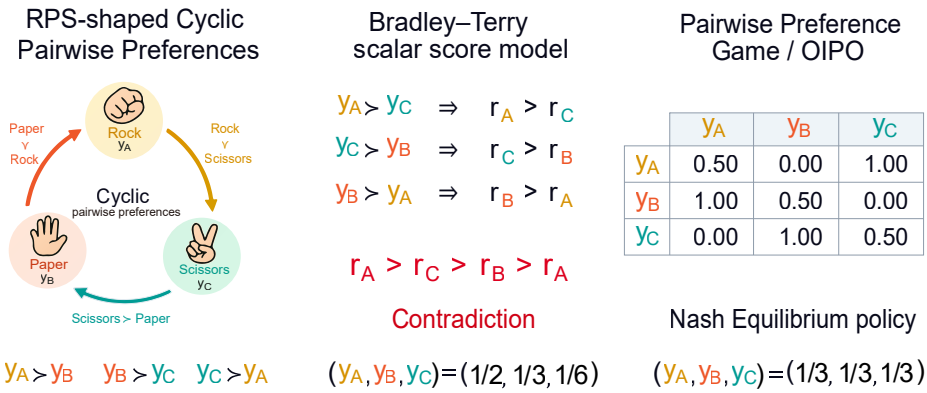

Figure 1: **Left:** A rock–paper–scissors example defines cyclic preferences. None of these three policies has a significant advantage over the others. **Middle:** Bradley–Terry-based methods require the preference pairs $r_A > r_C$, $r_C > r_B$, and $r_B > r_A$ simultaneously, raising a contradiction during learning. The final policy probability of rock, paper and scissors is $(1/2, 1/3, 1/6)$. **Right:** A preference game represents the pairwise preferences through the payoff matrix and admits a Nash equilibrium policy. The final policy probability is $(1/3, 1/3, 1/3)$.

incompatible constraints. For example, The pairwise preferences favor y_A over y_C, y_C over y_B, and y_B over y_A simultaneously. Under the BT assumption used by DPO, these three preference constraints cannot all be satisfied at the same time (Bradley & Terry, 1952; Rafailov et al., 2023). DPO must therefore trade off among the conflicting pairs, rather than faithfully representing the cyclic preference relation (Gheshlaghi Azar et al., 2024; Munos et al., 2024). Figure 1 illustrates this limitation with a rock–paper–scissors-shaped example.

A preference-game formulation provides an alternative way to model these relations. Reasoning candidates are treated as strategies, and the pairwise preference $P_x(y \succ y')$ directly defines the payoff between y and y', without requiring the comparison to be cast as a fixed chosen–rejected pair. Since these pairwise preferences are represented directly, they do not need to follow the Bradley–Terry structure, allowing cyclic relations to be represented without contradiction. For example, Nash Learning from Human Feedback (NLHF) targets a Kullback–Leibler (KL)-regularized Nash equilibrium of the resulting game (Munos et al., 2024; Calandriello et al., 2024). In the rock–paper–scissors example in Figure 1, the cyclic preferences are encoded directly in the payoff matrix, and the game admits a well-defined equilibrium distribution.

Prior work connects online preference optimization with Nash learning. In Online IPO and IPO-MD, each context is updated from a pair of sampled responses, so only one pairwise relation is observed per context at each collection step (Calandriello et al., 2024). Recent studies show that queried pair selection can materially affect feedback efficiency and downstream preference optimization (Muldrew et al., 2024; Xiao et al., 2025). Under a fixed comparison budget, this raises the question of how to allocate limited feedback across candidate pairs. Within a sampled pool, pair selection determines which relations receive feedback. More effective allocation can improve budget utilization, while changing the pair-sampling distribution may also alter the effective learning objective.

We therefore propose Coverage-Optimized Online Identity Preference Optimization (CO-OIPO), an implicit PRM framework for multi-hop knowledge-graph question answering that supports potentially intransitive preferences over reasoning trajectories and improves finite-budget preference-sampling efficiency through coverage-optimized pair selection. Candidate generation forms a finite pool from a current–reference proposal, while comparison allocation directs the limited feedback budget toward undercovered pairwise relations. Sampling corrections account for both collection stages so that the intended context-wise learning solution is preserved in expectation. At inference, the learned policy serves as an implicit PRM through its log-probability ratio to the reference policy, providing prefix-conditioned rewards for selecting the next reasoning step. Our contributions are threefold.

1. We propose CO-OIPO, an implicit process-reward learning framework that uses a preference-game objective in place of BT-based preference modeling. This formulation allows samples

with no clear preference distinction to receive comparable probability mass under the learned policy, reducing the risk of discarding potentially promising candidates and improving training stability. CO-OIPO further samples candidates from a current–reference mixture and prioritizes undercovered pairs under a fixed comparison budget.

2. We show that CO-OIPO can learn from preferences beyond the BT formulation, with its fixed point under exact online optimization corresponding to a KL-regularized Nash equilibrium. At this equilibrium, the PRM reward reflects how strongly a candidate is preferred against the equilibrium opponent, providing a meaningful process reward even for intransitive preferences that BT-based modeling cannot faithfully capture. We further show that CO-OIPO uses a limited preference budget more effectively: it preserves the intended per-question solution in expectation while allocating comparisons toward undercovered pairs to optimize coverage.
3. Experiments on WebQSP and Complex WebQuestions show the highest observed scores among the evaluated methods on three of the four reported dataset–metric combinations. Controlled preference-game experiments and fixed-pool analyses further examine equilibrium learning, pair selection coverage, and objective-estimation error.

2 Related Work

Knowledge-graph-enhanced multi-hop question answering. Retrieval-based systems such as RoG, GNN-RAG, and SubgraphRAG supply question-relevant graph evidence, while interactive approaches such as Graph-CoT and KG-RAG support repeated graph access during reasoning (Luo et al., 2024b; Mavromatis & Karypis, 2025; Li et al., 2025; Jin et al., 2024; Sanmartin, 2024). These approaches ground reasoning in structured evidence, but retrieving relevant evidence alone does not determine how competing reasoning trajectories should be compared. DPRM addresses graph–text coordination through mutually supporting KG-PRM and CoT-PRM evaluators (Wang et al., 2026). CO-OIPO complements evidence retrieval by focusing on the preference-learning objective and the allocation of limited trajectory comparisons. These benefits remain conditional on the availability of relevant evidence.

Process reward models. PRMs provide intermediate evaluation signals through explicit process supervision or trajectory-level feedback (Lightman et al., 2024; Wang et al., 2024; Luo et al., 2024a; Zhang et al., 2024). Explicit supervision offers direct step-level guidance but requires process-label construction. ImplicitPRM reduces this requirement by decomposing the policy–reference log ratio, while PRIME integrates implicit rewards into online reinforcement learning (Yuan et al., 2025; Cui et al., 2026). However, DPO-based implicit reward learning retains the BT restriction on preference representation. CO-OIPO retains log-ratio process scoring while grounding its learning target in a preference game.

Preference optimization and preference games. DPO simplifies policy learning through a BT preference likelihood, but its scalar-score representation cannot exactly capture cyclic preferences (Rafailov et al., 2023). IPO accommodates preference probabilities against a fixed opponent, while NLHF and online preference optimization connect learning to regularized Nash equilibria (Gheshlaghi Azar et al., 2024; Munos et al., 2024; Calandriello et al., 2024). Online IPO and IPO-MD observe one sampled pair per context at each collection step, leaving comparison coverage dependent on pair sampling. WPO addresses off-policy mismatch by reweighting existing preference pairs (Zhou et al., 2024). Reweighting can change the contribution of observed pairs but does not obtain feedback for unqueried relations. For online collection, changing pair probabilities also changes the uncorrected objective, so coverage and target preservation must be considered together. Building on these foundations, CO-OIPO optimizes which pairs receive feedback within a sampled pool and corrects both candidate generation and pair selection to preserve the intended context-wise minimizers in expectation.

3 Method

We present the preference-game learning target, coverage-aware sampling and corrected updates, process-reward-guided inference, and theoretical analysis in Sections 3.1–3.4, respectively.

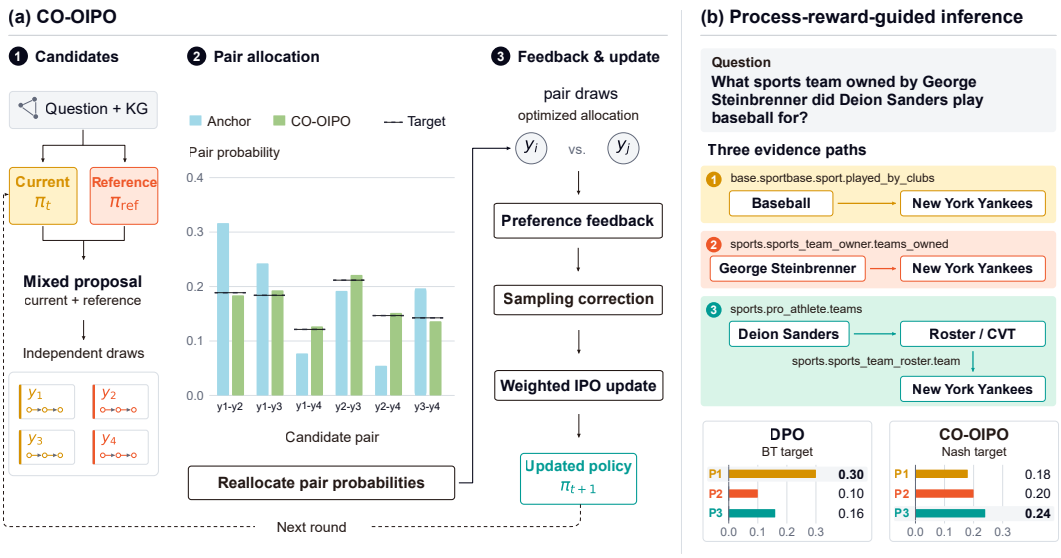

Figure 2: CO-OIPO training and inference. (a) Independent mixed-proposal draws form the candidate pool; the allocation plot is schematic. Pair allocation moves from anchor P_Y toward target $\bar{Q}_Y$; queried pairs use correction A_Y/M_Y^*. (b) An illustrative question and evidence paths. Inference uses the policy–reference log ratio on next-step candidates sharing a prefix.

3.1 Preference-Game Learning Target

CO-OIPO learns an implicit process reward from pairwise preferences over complete reasoning trajectories. We first define the preference-game target and round-wise objective; Section 3.2 then specifies the sampling procedure and correction.

Preference game. Let x contain a question and its retrieved graph evidence, and let $y = (z_1, \dots, z_H, \hat{a})$ denote a complete reasoning trajectory, where z_h is the reasoning step at position h and $\hat{a}$ is the final answer. A fixed comparison protocol provides a preference probability $P_x(y \succ y') \in [0,1]$, with $P_x(y \succ y') + P_x(y' \succ y) = 1$ and $1/2$ representing a neutral comparison. We use these pairwise preferences directly as the payoff of a two-player game (Munos et al., 2024; Calandriello et al., 2024). For two trajectory distributions π and ν, define

$$\mathcal{G}_x^\eta(\pi,\nu) = \mathbb{E}_{y\sim\pi,\, y'\sim\nu}[P_x(y \succ y')] - \eta\,\mathrm{KL}(\pi\|\pi_{\mathrm{ref}}) + \eta\,\mathrm{KL}(\nu\|\pi_{\mathrm{ref}}), \qquad \max_\pi \min_\nu \mathcal{G}_x^\eta(\pi,\nu). \quad (1)$$

All policies condition on x, with fixed reference π_{ref} and regularization strength $\eta > 0$. At a symmetric Nash equilibrium, neither player can improve its regularized payoff by unilateral deviation. Using P_x directly avoids requiring a Bradley–Terry representation.

Round-wise target and CO-OIPO objective. Following the geometric-mixture construction used in NLHF (Munos et al., 2024), the current policy π_t and reference define

$$g_t(y \mid x) = \pi_t(y \mid x)^{1-\alpha}\pi_{\mathrm{ref}}(y \mid x)^\alpha, \qquad q_t(y \mid x) = \frac{g_t(y \mid x)}{Z_t(x)}, \qquad Z_t(x) = \sum_y g_t(y \mid x). \quad (2)$$

Here g_t is the unnormalized target mass, q_t is the target distribution, and $\alpha \in [0,1)$ controls the reference contribution. This target is fixed within a round and refreshed after each update.

CO-OIPO generates a finite candidate pool and queries only a limited number of pairs. For a realized pool $Y(x)$, let $A_{Y(x)}(e)$ be the target mass of pair e and $M^*_{Y(x)}(e)$ its query probability. Define $f_\theta(y \mid x) = \log[\pi_\theta(y \mid x)/\pi_{\mathrm{ref}}(y \mid x)]$ and the pair difference $h_\theta(x; y, y') = f_\theta(y \mid x) - f_\theta(y' \mid x)$. For observed feedback $v \in \{0, 1/2, 1\}$ (right win, tie, or left win), the IPO pair loss is

$$\widehat{\ell}_\theta(x; y, y'; v) = v(h_\theta - m)^2 + (1-v)(h_\theta + m)^2, \qquad m = \frac{1}{2\tau}, \quad \tau > 0. \quad (3)$$

This uses the IPO target margin (Gheshlaghi Azar et al., 2024); its expectation is denoted by ℓ_θ. Given a context batch $\mathcal{B}_t$ and B queried pairs per active context, CO-OIPO optimizes

$$\widehat{\mathcal{L}}_t^{\mathrm{CO}}(\theta) = \frac{1}{|\mathcal{B}_t|} \sum_{x \in \mathcal{B}_t} \frac{1}{B} \sum_{b=1}^{B} \frac{A_{Y(x)}(e_b)}{M^*_{Y(x)}(e_b)} \widehat{\ell}_\theta(x, e_b). \tag{4}$$

For each queried pair, $A_Y(e)$ specifies its target contribution, while $M^*_Y(e)$ is its query probability. The ratio corrects their mismatch and accounts for both candidate generation and pair selection. Only π_θ receives gradients.

Implicit process reward. The learned policy provides a process reward through its log-probability ratio to the reference, without an additional reward head. For a candidate step z_h conditioned on the selected prefix $z_{<h}$, define

$$r_\theta(z_h \mid x, z_{<h}) = \beta \log \frac{\pi_\theta(z_h \mid x, z_{<h})}{\pi_{\mathrm{ref}}(z_h \mid x, z_{<h})}, \qquad \beta > 0. \tag{5}$$

This reward sums the step's token log ratios and is positive when the learned policy assigns that complete step more probability than the reference does. After T rounds, the inference reward is $s_T = r_T$.

3.2 Coverage-Aware Sampling and Corrected Updates

We construct A_Y and M^*_Y in three stages. Coverage refers to pair-query probabilities relative to the target pair distribution within the realized pool.

Stage 1: mixed-proposal candidate generation. For each of $K \geq 2$ candidate positions, independently sample a complete trajectory from

$$S_t(y \mid x) = (1 - \lambda_t)\pi_t(y \mid x) + \lambda_t \pi_{\mathrm{ref}}(y \mid x), \qquad a_i = \frac{g_t(y_i \mid x)}{S_t(y_i \mid x)}. \tag{6}$$

The coefficient $\lambda_t \in [0, 1]$ controls candidate collection, with $\lambda_t = 1/2$ for QA; α defines the learning target. The weight a_i corrects the proposal toward g_t. Retain repeated trajectories as separate pool positions. For an unordered index pair $e = (i, j)$, $i < j$, define

$$A_Y(i, j) = \frac{2a_i a_j}{K(K-1)} \mathbf{1}\{y_i \neq y_j\}, \qquad \kappa_Y = \sum_e A_Y(e), \qquad \bar{Q}_Y(e) = \frac{A_Y(e)}{\kappa_Y}. \tag{7}$$

The factor $2/[K(K-1)]$ averages over unordered candidate pairs, while the indicator excludes identical-trajectory comparisons. We retain A_Y in the loss and use its normalized distribution $\bar{Q}_Y$ for pair allocation. A pool with $\kappa_Y = 0$ contributes zero and remains in the context-batch denominator.

Stage 2: coverage-optimized pair allocation. Keep the candidate pool fixed and define the current-policy anchor allocation

$$b_i = \frac{\pi_t(y_i \mid x)}{S_t(y_i \mid x)}, \qquad P_Y(i, j) = \frac{b_i b_j \mathbf{1}\{y_i \neq y_j\}}{\sum_{u<v} b_u b_v \mathbf{1}\{y_u \neq y_v\}}. \tag{8}$$

CO-OIPO reallocates pair probability from the anchor P_Y toward the target $\bar{Q}_Y$, subject to

$$(1 - \gamma_t)P_Y(e) \leq M(e) \leq P_Y(e) + \gamma_t \bar{Q}_Y(e), \qquad \gamma_t \in [0, 1], \qquad \sum_e M(e) = 1. \tag{9}$$

The lower bound retains at least a $(1 - \gamma_t)$ fraction of each pair's anchor probability, while the upper bound limits its increase according to $\bar{Q}_Y(e)$. Thus larger γ_t permits stronger movement toward the target. Within these bounds, CO-OIPO uses the water-filling allocation

$$M^*_Y(e) = \mathrm{clip}\left(\frac{\bar{Q}_Y(e)}{c_Y}, (1 - \gamma_t)P_Y(e), P_Y(e) + \gamma_t \bar{Q}_Y(e)\right), \qquad \sum_e M^*_Y(e) = 1. \tag{10}$$

Here $\mathrm{clip}(v, l, u)$ restricts v to $[l, u]$, and $c_Y > 0$ is chosen by a one-dimensional monotone search. This gives a water-filling allocation: unconstrained pairs receive probability proportional to $\bar{Q}_Y$, while the others remain at their lower or upper bounds. At $\gamma_t = 0$, $M^*_Y = P_Y$; at $\gamma_t = 1$, $M^*_Y = \bar{Q}_Y$.

Stage 3: preference feedback and corrected update. For each active pool, independently sample B pairs with replacement from M_Y^* and obtain feedback from the fixed comparison protocol. Keep the candidate pool, query probabilities, correction weights, and feedback fixed while optimizing the round. The correction weight A_Y/M_Y is applied directly to each sampled pair loss, without clipping or batch-level normalization. Starting from $\pi_0 = \pi_{\text{ref}}$, the update produces π_{t+1} for the next round; after T rounds, use π_T for inference.

3.3 Process-Reward-Guided Inference

Given the selected prefix $z_{<h}$, a proposal language model generates next-step candidates, and CO-OIPO selects the largest process reward:

$$\mathcal{C}_h = \{z_h^{(1)}, \ldots, z_h^{(N)}\}, \qquad \hat{z}_h = \arg\max_{z \in \mathcal{C}_h} s_T(z \mid x, z_{<h}), \quad s_T = \beta \log \frac{\pi_T(z \mid x, z_{<h})}{\pi_{\text{ref}}(z \mid x, z_{<h})}. \tag{11}$$

Each candidate contains a reasoning step and its evidence transition. Its reward sums the token log ratios directly. Append $\hat{z}_h$ to the retained prefix and generate the next candidate set from that updated prefix. Repeat until completion or the maximum number of reasoning steps, then generate the final answer from the selected sequence. The QA experiments use $N = 8$ and at most four steps.

3.4 Theoretical Analysis

The analysis assumes a finite space of complete trajectories, positive reference and collection probabilities, reciprocal expected feedback, and the full correction in Eq. (4). Appendix B gives detailed assumptions and proofs.

Finite-budget target preservation. Let $\ell_\theta(x, y, y')$ be the expectation of Eq. (3) under feedback satisfying $\mathbb{E}[v \mid x, y, y'] = P_x(y \succ y')$. The round-wise population objective is $\mathcal{L}_{t,x}(\theta) = \mathbb{E}_{y,y' \overset{\text{iid}}{\sim} q_t}[\ell_\theta(x, y, y')]$. For a context x, let $\widehat{R}_{t,x}$ be the inner average over B pairs in Eq. (4).

Theorem 3.1 (Finite-budget target preservation)**.** *For every finite $K \geq 2$ and $B \geq 1$, the expectation over candidate generation, pair queries, and feedback satisfies*

$$\mathbb{E}[\widehat{R}_{t,x}(\theta)] = Z_t(x)^2 \mathcal{L}_{t,x}(\theta) - c_{t,x}, \qquad c_{t,x} = \frac{1}{4\tau^2} \sum_y g_t(y \mid x)^2. \tag{12}$$

Consequently, the expected corrected estimator and $\mathcal{L}_{t,x}$ have the same context-wise minimizers.

The subtracted constant accounts for identical-trajectory pairs. Across contexts, the expected objective weights $\mathcal{L}_{t,x}$ by $Z_t(x)^2$.

Nash correspondence and process-reward interpretation. The context-wise solution connects the corrected update to the NLHF preference game through the online IPO optimizer (Munos et al., 2024; Calandriello et al., 2024).

Corollary 3.2 (Nash correspondence and implicit process reward)**.** *If the policy class represents every positive conditional distribution and each round is optimized exactly for each context, a fixed point π_α^* induces*

$$q_\alpha^*(y \mid x) \propto \big(\pi_\alpha^*(y \mid x)\big)^{1-\alpha} \pi_{\text{ref}}(y \mid x)^\alpha, \qquad \eta = \frac{\tau}{1-\alpha}, \tag{13}$$

which is the symmetric Nash equilibrium of Eq. (1)*. Define $w_x(y) = \mathbb{E}_{y' \sim q_\alpha^*(\cdot|x)}[P_x(y \succ y')]$. Then $\log[\pi_\alpha^*(y \mid x)/\pi_{\text{ref}}(y \mid x)] = w_x(y)/\tau - C_\alpha(x)$, where $C_\alpha(x)$ is independent of y. For a valid prefix s and complete continuation u, define*

$$V_x(s) = \tau \log \mathbb{E}_{u \sim \pi_{\text{ref}}(\cdot|x,s)} \left[\exp\left(\frac{w_x(s \oplus u)}{\tau} \right) \right]. \tag{14}$$

For a next-step completion z ending at a valid step boundary,

$$r_\alpha^*(z \mid x, s) = \beta \log \frac{\pi_\alpha^*(z \mid x, s)}{\pi_{\text{ref}}(z \mid x, s)} = \frac{\beta}{\tau} \big[V_x(s \oplus z) - V_x(s) \big]. \tag{15}$$

Thus a next step contributes the increment in a reference-regularized continuation value. Under a common prefix, Eq. (11) selects the largest continuation value. The equilibrium is q_α^*, the geometric transform of the output policy π_α^*.

Coverage-optimal pair allocation. For an active pool, define

$$\mathfrak{C}_Y(M) = \max_{e:\bar{Q}_Y(e)>0} \frac{\bar{Q}_Y(e)}{M(e)}, \qquad \mathfrak{J}_Y(M) = \sum_{e:\bar{Q}_Y(e)>0} \frac{\bar{Q}_Y(e)^2}{M(e)}. \tag{16}$$

Here $\mathfrak{C}_Y$ measures worst relative pair undercoverage and $\mathfrak{J}_Y$ is the second moment of normalized pair importance weights.

Proposition 3.3 (Coverage-optimal pair allocation)**.** *On an active pool, among pair distributions satisfying Eq.* (9)*, the allocation M_Y^* satisfies*

$$M_Y^* \in \arg\min_M \mathfrak{C}_Y(M) \;\cap\; \arg\min_M \mathfrak{J}_Y(M). \tag{17}$$

In particular, $\mathfrak{C}_Y(M_Y^*) \le \mathfrak{C}_Y(P_Y)$ and $\mathfrak{J}_Y(M_Y^*) \le \mathfrak{J}_Y(P_Y)$.

These loss-independent criteria determine a common conditional estimation bound; actual estimation error also depends on per-pair losses. Appendix B.5 gives finite-pool and pair-estimation bounds and an anchor-relative gradient-second-moment bound.

4 Experiments

We evaluate CO-OIPO along four dimensions: (1) end-to-end answer quality on multi-hop QA, (2) the contribution of proposal mixing and pair allocation, (3) equilibrium learning in controlled preference games, and (4) pair-sampling efficiency on fixed candidate pools.

4.1 Experimental Setup and Answer Quality

Benchmarks and metrics. We evaluate on the full official test sets of WebQSP and Complex WebQuestions (CWQ), two multi-hop QA benchmarks built on Freebase (Yih et al., 2016; Talmor & Berant, 2018; Bollacker et al., 2008). We report **Hit**, the percentage of questions whose predicted answer set contains at least one gold answer, and **F1**, the harmonic mean of answer-set precision and recall. Higher is better for both metrics.

Models and training. All preference-optimization methods use Qwen3-8B (Yang et al., 2025) and are initialized directly from the same base-model checkpoint. A frozen copy of the base model serves as the reference policy. DPO, IPO, IPO-MD, and CO-OIPO are optimized with AdamW (Loshchilov & Hutter, 2019) and trained separately on the same questions within each dataset. CO-OIPO runs three online rounds, generating four trajectories and sampling two pairs per question in each round. A DeepSeek-V4-Flash judge provides pairwise preference feedback during QA training. At inference, a Qwen3-8B instruction model proposes next-step candidates, and the trained PRM selects the next step. Additional data and implementation details are provided in Appendix C.

Baselines. Table 1 compares LLM-only systems, KG-enhanced methods, outcome reward models, and process reward models. We use published results when available and run the remaining evaluations. The DPO, IPO, IPO-MD, and CO-OIPO rows share the base initialization and dataset-specific training questions (Rafailov et al., 2023; Gheshlaghi Azar et al., 2024; Calandriello et al., 2024).

Answer quality. CO-OIPO leads on three of four metrics (Table 1), improving WebQSP Hit over DCTR by 0.24 points and WebQSP F1 and CWQ Hit over DPRM by 1.18 and 1.10 points, respectively. Its CWQ F1 is 49.67, below DCTR (50.12) and DPRM (59.88).

Among the preference-learning methods sharing the same initialization and training questions, CO-OIPO improves over IPO-MD by 7.58, 4.11, 5.60, and 4.07 percentage points on WebQSP Hit, WebQSP F1, CWQ Hit, and CWQ F1, respectively.

Table 1: **Answer quality on the full WebQSP and CWQ test sets** (%, higher is better). The final four rows share the Qwen3-8B base initialization and dataset-specific training questions. Bold indicates the highest observed result in each column.

Category	Method	WebQSP		CWQ	
		Hit ↑	F1 ↑	Hit ↑	F1 ↑
LLM-only	Qwen3-8B (Yang et al., 2025)	62.84	35.36	37.75	25.64
	Llama-3.1-8B (Grattafiori et al., 2024)	55.50	34.80	28.10	22.40
KG-enhanced	SubgraphRAG (Li et al., 2025)	86.61	70.57	56.98	47.16
	DCTR (Cao et al., 2026)	88.05	72.13	59.72	50.12
	ParallaxRAG (Liu et al., 2026a)	86.85	71.73	58.41	48.33
Outcome RM	Skywork-Reward-V2 (Liu et al., 2026c)	63.82	45.70	32.80	28.66
	DeepSeek-GRM (Liu et al., 2025)	61.61	43.90	30.50	26.26
	RM-R1 (Chen et al., 2026)	69.47	50.19	35.17	30.61
Process RM	ThinkPRM (Khalifa et al., 2026)	64.68	46.00	31.89	27.86
	VersaPRM (Zeng et al., 2025)	65.79	46.94	31.80	27.90
	PRInTS (Lee et al., 2026)	65.42	46.80	32.85	28.86
	DPRM (Wang et al., 2026)	87.15	72.28	73.01	**59.88**
	DPO PRM (Rafailov et al., 2023)	72.42	54.67	56.50	35.34
	IPO PRM (Gheshlaghi Azar et al., 2024)	76.17	65.23	57.38	34.15
	IPO-MD PRM (Calandriello et al., 2024)	80.71	69.35	68.51	45.60
	CO-OIPO	**88.29**	**73.46**	**74.11**	49.67

4.2 Ablation of Proposal Mixing and Pair Allocation

We study the contributions of the two sampling components of CO-OIPO: proposal mixing and pair allocation. The ablation uses the asymmetric four-action preference game with Qwen3-8B and LoRA (Hu et al., 2022). All variants share the same initialization $\pi_0 = (0.075, 0.75, 0.075, 0.1)$, reference policy $\rho = (0.3, 0.3, 0.3, 0.1)$, learning target, and training budget. Proposal mixing uses $\lambda = 0.75$ when enabled and $\lambda = 0$ otherwise; pair allocation uses $\gamma = 0.8$ when enabled and the anchor allocation otherwise. Importance corrections are retained in all variants.

We report output-policy **exploitability** $E(\pi) = \max_a (P\pi)_a - 0.5$ and the **regularized Nash gap** $\mathrm{Gap}_\eta(q) = \max_u G_\eta(u, q) - \min_v G_\eta(q, v)$, where P is the four-action preference matrix and G_η is Eq. (1) with reference ρ. All controlled games use $q \propto \pi^{0.5}\rho^{0.5}$ and $\eta = 0.2$; lower is better for both metrics. Multi-seed standard deviations and individual runs are provided in Appendix E.5.

Table 2: **Ablation of the two sampling components in CO-OIPO.** We evaluate *proposal mixing*, which combines current and reference policies, and *pair allocation*, which improves coverage of the target pair distribution, on the asymmetric four-action game. Lower is better for exploitability $E(\pi)$ and regularized Nash gap $\mathrm{Gap}_\eta(q)$.

Proposal mixing	Pair allocation	$E(\pi) \downarrow$	$\mathrm{Gap}_\eta(q) \downarrow$
Off	Off	0.032820	0.003077
On	Off	0.036656	0.002920
Off	On	0.030022	0.003253
On	On	**0.027967**	**0.000748**

The joint configuration reduces both exploitability and the regularized Nash gap relative to the all-off setting. Neither component alone improves both metrics.

4.3 Learning the Preference-Game Target

We compare offline DPO and CO-OIPO on symmetric and asymmetric four-action games with cyclic preferences among A, B, C and a dominated action N. Both methods share the calibrated Qwen3-8B initialization, reference policy, and training budget. DPO uses a fixed offline preference dataset,

whereas CO-OIPO collects pairs online from the evolving policy and optimizes a preference-game objective. We use the exploitability and regularized Nash-gap metrics from Section 4.2; settings and multi-seed results are provided in Appendix E.

Table 3: **Equilibrium learning in controlled preference games.** Offline DPO and CO-OIPO are compared on symmetric and asymmetric four-action games with cyclic preferences and a dominated action. Lower is better for exploitability $E(\pi)$ and regularized Nash gap $\mathrm{Gap}_\eta(q)$.

Game	Method	$E(\pi) \downarrow$	$\mathrm{Gap}_\eta(q) \downarrow$
Symmetric	DPO	0.06346	0.00702
	CO-OIPO	**0.03975**	**0.00482**
Asymmetric	DPO	0.08231	0.01353
	CO-OIPO	**0.04484**	**0.00350**

CO-OIPO achieves lower mean exploitability and regularized Nash gap than offline DPO in both controlled games, with larger improvements in the asymmetric setting (Table 3). These results support closer tracking of the preference-game target under the evaluated settings. Appendix E provides training details, parameter selection, and multi-seed results.

4.4 Pair-Sampling Coverage and Loss Estimation

We isolate pair allocation on frozen WebQSP pools with four candidates and two pair draws per pool, holding candidates, target weights, preference labels, and checkpoints fixed. Anchor, Linear, and CO-OIPO sample from P_Y, $(P_Y + \bar{Q}_Y)/2$, and M_Y^*, respectively, under identical probability constraints ($\gamma = 0.5$) and importance correction.

We evaluate **worst undercoverage** $\mathfrak{C}_Y - 1$, **allocation mismatch** $\mathfrak{J}_Y - 1 = \chi^2(\bar{Q}_Y \| M)$, and the exact conditional mean-squared error $\mathrm{MSE}_Y = \mathbb{E}[(\widehat{U}_Y - U_Y)^2 \mid Y]$ of the corrected loss under two independent pair draws. Both the coverage statistics and loss-estimation errors are computed over the same 4,750 active pools. Lower is better for all metrics; zero is ideal for the two coverage criteria.

Table 4: **Comparison of pair allocations and corrected-loss estimation on frozen WebQSP candidate pools.** $\mathfrak{C}_Y - 1$ measures the worst-case allocation shortfall relative to the target pair distribution, $\mathfrak{J}_Y - 1 = \chi^2(\bar{Q}_Y \| M)$ measures overall allocation mismatch, and mean/median MSE summarize the conditional error of the corrected loss estimator. Lower is better for all metrics.

Sampler	$\mathfrak{C}_Y - 1 \downarrow$	$\mathfrak{J}_Y - 1 \downarrow$	Mean MSE $\downarrow$	Median MSE $\downarrow$
Anchor	0.3297	0.0572	0.439862	0.015215
Linear	0.1134	0.0118	0.250842	0.012568
CO-OIPO	**0.0084**	**0.0008**	**0.224850**	**0.011748**

CO-OIPO achieves the lowest worst undercoverage, allocation mismatch, mean loss MSE, and median loss MSE among the three samplers (Table 4). Relative to Anchor, it reduces mean MSE by 48.9%; relative to Linear, the reduction is 10.4%. Appendix D provides additional pool-level details and the exact MSE calculation under the same 4,750-pool evaluation set.

5 Conclusion

We introduced CO-OIPO, which connects the NLHF preference-game target to implicit process rewards through mixed-proposal candidate generation, optimized within-pool pair allocation, and importance corrections at both sampling stages. Under the stated assumptions, the corrected estimator preserves exact context-wise minimizers and the associated Nash fixed-point correspondence. CO-OIPO achieves the highest performance on most metrics. Controlled cyclic-game experiments show lower mean regularized Nash gaps than offline DPO, with variation across seeds. On frozen WebQSP pools, CO-OIPO reduces mean loss-estimation error relative to Anchor and Linear.

AI USE STATEMENT

In this work, we used generative AI tools for tasks with required disclosure, including generating synthetic datasets used in our experiments and assisting with proving mathematical claims. For the latter, generative AI tools were used to help develop, check, and refine intermediate mathematical reasoning and proofs; the final mathematical claims, derivations, and proofs presented in the paper were reviewed and verified by the authors.

Additionally, we used generative AI tools for tasks with recommended disclosure. Specifically, they were used to aid and polish the writing of the manuscript, to draft portions of the paper, and for retrieval and discovery of potentially relevant prior work. All cited references identified with AI assistance were subsequently checked by the authors.

We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, results, references, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The QA experiments use public knowledge-graph question-answering benchmarks; the four-action cyclic-game experiments use synthetic feedback. Automated preference judgments can reflect model biases and errors; the results should not be interpreted as establishing factual reliability outside the evaluated settings. Data and model use remain subject to the original licenses.

REPRODUCIBILITY STATEMENT

Section 4.1 and Appendix C specify the data splits, initialization, optimizer, and principal sampling budgets. The appendices provide estimator details, theoretical assumptions and proofs, training protocols, and fixed-pool evaluation procedures. Reported sampling and estimation analyses specify their reference targets and evaluation units; the controlled-game appendix reports all three seeds. The anonymous code repository is available at `https://anonymous.4open.science/r/co-oipo-82FB`.

REFERENCES

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pp. 1247–1250, 2008.

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. ISSN 00063444, 14643510. URL http://www.jstor.org/stable/2334029.

Daniele Calandriello, Zhaohan Daniel Guo, Remi Munos, Mark Rowland, Yunhao Tang, Bernardo Avila Pires, Pierre Harvey Richemond, Charline Le Lan, Michal Valko, Tianqi Liu, Rishabh Joshi, Zeyu Zheng, and Bilal Piot. Human alignment of large language models through online preference optimisation. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5409–5435. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/calandriello24a.html.

Yukun Cao, Zirui Xu, Dongyang Li, Zhihao Guo, Luobin Huang, and Lisheng Wang. DCTR: Dual-constraint subgraph optimization for knowledge graph-based retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 30156–30164, 2026.

Xiusi Chen, Gaotang Li, Ziqi Wang, Bowen Jin, Cheng Qian, Yu Wang, Hongru WANG, Yu Zhang, Denghui Zhang, Tong Zhang, Hanghang Tong, and Heng Ji. RM-R1: Reward modeling as reasoning. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 88313–88342, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/8e3b8de251afd887fb4589c1e3a3c793-Paper-Conference.pdf.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, Kaiyan Zhang, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, Hao Peng, Yu Cheng, Zhiyuan Liu, Maosong Sun, Bowen Zhou, and Ning Ding. Process reinforcement through implicit rewards. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=9SkkifLopZ.

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 Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4447–4455. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/gheshlaghi-azar24a.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. In *Breakthroughs in statistics: Foundations and basic theory*, pp. 308–334. Springer, 1992.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Bowen Jin, Chulin Xie, Jiawei Zhang, Kashob Kumar Roy, Yu Zhang, Zheng Li, Ruirui Li, Xianfeng Tang, Suhang Wang, Yu Meng, and Jiawei Han. Graph chain-of-thought: Augmenting large language models by reasoning on graphs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 163–184, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.11. URL https://aclanthology.org/2024.findings-acl.11/.

Muhammad Khalifa, Rishabh Agarwal, Lajanugen Logeswaran, Jaekyeom Kim, Hao Peng, Moontae Lee, Honglak Lee, and Lu Wang. Process reward models that think. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=FPVCb0WMuN. J2C Certification.

Jaewoo Lee, Archiki Prasad, Justin Chen, Zaid Khan, Elias Stengel-Eskin, and Mohit Bansal. PRInTS: Reward modeling for long-horizon information seeking. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 34120–34138, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1574. URL https://aclanthology.org/2026.acl-long.1574/.

Mufei Li, Siqi Miao, and Pan Li. Simple is effective: The roles of graphs and large language models in knowledge-graph-based retrieval-augmented generation. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 6061–6089, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/11e1900e680f5fe1893a8e27362dbe2c-Paper-Conference.pdf.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/aca97732e30bcf1303bc22ac3924fd16-Paper-Conference.pdf.

Jinliang Liu, Jiale Bai, and Shaoning Zeng. Think parallax: Solving multi-hop problems via multi-view knowledge-graph-based retrieval-augmented generation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26637–26655, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1226. URL https://aclanthology.org/2026.acl-long.1226/.

Xiaoqian Liu, Ke Wang, Yuchuan Wu, Fei Huang, Yongbin Li, Jianbin Jiao, and Junge Zhang. Agentic reinforcement learning with implicit step rewards. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 129271–129291, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/d1e6739f319be4ba8e748cc05f23bba1-Paper-Conference.pdf.

Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, and Yang Liu. Skywork-Reward-V2: Scaling preference data curation via human-AI synergy. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 133805–133838, 2026c. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/d8836d0294651a867c2ca51012923997-Paper-Conference.pdf.

Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. Inference-time scaling for generalist reward modeling, 2025. URL https://arxiv.org/abs/2504.02495.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, and Abhinav Rastogi. Improve mathematical reasoning in language models by automated process supervision, 2024a. URL https://arxiv.org/abs/2406.06592.

Linhao Luo, Yuan-Fang Li, Reza Haffari, and Shirui Pan. Reasoning on graphs: Faithful and interpretable large language model reasoning. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 14400–14423, 2024b. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/3e2aeb66481dd63a32421bf032b70384-Paper-Conference.pdf.

Costas Mavromatis and George Karypis. GNN-RAG: Graph neural retrieval for efficient large language model reasoning on knowledge graphs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 16682–16699, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.856. URL https://aclanthology.org/2025.findings-acl.856/.

William Muldrew, Peter Hayes, Mingtian Zhang, and David Barber. Active preference learning for large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36577–36590. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/muldrew24a.html.

Remi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Côme Fiegel, Andrea Michi, Marco Selvi, Sertan Girgin, Nikola Momchev, Olivier Bachem, Daniel J Mankowitz, Doina Precup, and Bilal Piot. Nash learning from human feedback. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36743–36768. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/munos24a.html.

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. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741. Curran Associates, Inc., 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf.

Diego Sanmartin. KG-RAG: Bridging the gap between knowledge and creativity, 2024. URL https://arxiv.org/abs/2405.12035.

Alon Talmor and Jonathan Berant. The web as a knowledge-base for answering complex questions. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 641–651, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1059. URL https://aclanthology.org/N18-1059/.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.510. URL https://aclanthology.org/2024.acl-long.510/.

Xinyi Wang, Yiping Song, Zhiliang Tian, Bo Liu, Tingjin Luo, and Minlie Huang. DPRM: A dual implicit process reward model in multi-hop question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33683–33691, 2026.

Yao Xiao, Hai Ye, Linyao Chen, Hwee Tou Ng, Lidong Bing, Xiaoli Li, and Roy Ka-Wei Lee. Finding the sweet spot: Preference data construction for scaling preference optimization. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the*

63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 12538–12552, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.615. URL https://aclanthology.org/2025.acl-long.615/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Wen-tau Yih, Matthew Richardson, Chris Meek, Ming-Wei Chang, and Jina Suh. The value of semantic parse labeling for knowledge base question answering. In Katrin Erk and Noah A. Smith (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 201–206, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-2033. URL https://aclanthology.org/P16-2033/.

Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. Free process rewards without process labels. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 73511–73525. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/yuan25c.html.

Thomas Zeng, Shuibai Zhang, Shutong Wu, Christian Classen, Daewon Chae, Ethan Ewer, Minjae Lee, Heeju Kim, Wonjun Kang, Jackson Kunde, Ying Fan, Jungtaek Kim, Hyung Il Koo, Kannan Ramchandran, Dimitris Papailiopoulos, and Kangwook Lee. VersaPRM: Multi-domain process reward model via synthetic reasoning data. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 74197–74239. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/zeng25h.html.

Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. Rest-mcts$*$: LLM self-training via process reward guided tree search. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 64735–64772. Curran Associates, Inc., 2024. doi: 10.52202/079017-2066. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/76ec4dc30e9faaf0e4b6093eaa377218-Paper-Conference.pdf.

Wenxuan Zhou, Ravi Agrawal, Shujian Zhang, Sathish Reddy Indurthi, Sanqiang Zhao, Kaiqiang Song, Silei Xu, and Chenguang Zhu. WPO: Enhancing RLHF with weighted preference optimization. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8328–8340, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.475. URL https://aclanthology.org/2024.emnlp-main.475/.

A Estimator and Implementation Details

A.1 Notation and collection procedure

Table 5: Notation used throughout the manuscript.

Symbol	Meaning
x, y, s, z	Context, complete trajectory, valid prefix, and next-step completion.
ρ, π_t, π_θ	Fixed reference, frozen collection policy, and trainable policy.
$P_x(y \succ y')$, v	Expected reciprocal preference and observed pair feedback.
g_t, Z_t, q_t	Geometric target mass, its total mass, and target distribution.
$S_{t,\lambda}$	Arithmetic proposal over complete trajectories.
Y, $\mathcal{E}_Y$, K, B	Candidate pool, valid index pairs, pool size, and pair-draw budget.
a_i, b_i	Geometric-target and current-policy proposal corrections.
A_Y, κ_Y, $\bar{Q}_Y$	Unnormalized pair mass, its sum, and target pair distribution.
P_Y, M_Y^*	Anchor and optimized within-pool pair allocations.
α, λ, γ	Geometric mixing, proposal mixing, and pair redistribution parameters.
τ, η, β	IPO scale, game regularization, and process-reward scale.

The symbols P_x and P_Y have distinct roles: P_x is the preference kernel, whereas P_Y is a distribution over index pairs in a particular pool. The target distribution q_t is geometric; the proposal $S_{t,\lambda}$ is arithmetic. Their probability normalizers are part of the sampling construction.

For each context and round, independently choose the current policy with probability $1-\lambda$ or the reference with probability λ, then sample a complete trajectory from the chosen component. Repeat this procedure K times. Thus each position has proposal probability $S_{t,\lambda}(y \mid x)$ in Eq. (6). Compute each complete-trajectory log probability with its termination convention, and evaluate the mixture probability by log-sum-exp. Selecting one component per trajectory is different from mixing the two token distributions at every token.

Keep all pool positions, including duplicates. Exclude only index pairs whose trajectory strings are identical. For an active pool, form A_Y, $\bar{Q}_Y$, and P_Y using Eq. (7); solve Eq. (10); and sample B pairs independently with replacement. Present the pair in a fixed canonical order, attach its observed label, and retain repeated pair draws as repeated observations. Freeze all collection quantities during the round's optimization. A zero-contribution pool requires no feedback and contributes zero to the original context average.

A.2 Allocation solver and endpoints

For $0 < \gamma < 1$, the function

$$F(c) = \sum_e \operatorname{clip}\big(\bar{Q}_Y(e)/c, L_Y(e), U_Y(e)\big)$$

is continuous and nonincreasing, with limits $F(0^+) = 1+\gamma$ and $F(+\infty) = 1-\gamma$. A root of $F(c) = 1$ exists and can be found by bisection after bracketing. The allocation is unique even if several values of c produce it. Use the exact endpoint distributions P_Y and $\bar{Q}_Y$ when γ is zero or one. The sums in Eq. (7) and the simplex constraint are probability normalizations, not transformations of the inference reward.

A.3 Feedback loss and zero-contribution pools

For expected preference $p = P_x(y \succ y')$, the loss used in the analysis is

$$\ell_\theta = p(h_\theta - m)^2 + (1-p)(h_\theta + m)^2 = h_\theta^2 - \frac{2p-1}{\tau} h_\theta + \frac{1}{4\tau^2}. \tag{A.1}$$

The hard-label loss equals $(h_\theta - (2v-1)/(2\tau))^2$ for $v \in \{0,1\}$. For a tie $v = 1/2$, Eq. (3) is $h_\theta^2 + 1/(4\tau^2)$. Removing this constant leaves the same gradient but changes the loss values used in MSE diagnostics; the reported estimator is defined by Eq. (3).

Identical trajectories have $h_\theta = 0$ and constant expected loss $1/(4\tau^2)$. Omitting these comparisons removes a parameter-independent term, which is accounted for in Theorem 3.1. In contrast, discarding a context with no valid pair and averaging only over the remaining contexts changes the context measure. Zero pools therefore remain in the denominator. When an experiment instead continues collection to obtain a fixed number of informative updates, both zero-pool attempts and candidate draws must be counted separately, as in Appendix E.5.

A.4 Process reward and inference

For step tokens $z = (z_1, \ldots, z_m)$, the reward is

$$r_\theta(z \mid x, s) = \beta \sum_{j=1}^{m} \log \frac{\pi_\theta(z_j \mid x, s, z_{<j})}{\rho(z_j \mid x, s, z_{<j})}. \tag{A.2}$$

The sum includes the tokens specifying the end of the step under the chosen segmentation. At the final step, the complete trajectory also includes the answer and termination. The reward in Eq. (A.2) is used directly to select one next-step candidate from the common prefix. Multiplication by a fixed positive β does not change that selection. Only the selected prefix is extended; the described procedure does not maintain a beam of alternative prefixes.

A.5 Implementation conditions for the expectation identity

The proposal denominator must be the actual probability of the generated complete trajectory. Temperature, truncation, retries, and fallback branches alter this probability. Full target support must be retained, or the trajectory space and all target policies must be consistently restricted. The exact estimator uses independent candidate positions, retained multiplicities, with-replacement pair draws, fixed expected feedback, and unmodified weights A_Y/M_Y^*. Reciprocal expected feedback is a property of the comparison protocol required by the preference-game analysis; a single presentation order does not by itself establish this property for a language-model judge.

If pairs are sampled without replacement, Eq. (4) is not the appropriate estimator. For a sampled set D with known positive inclusion probabilities ι_e, the conditional estimator is $\sum_{e \in D} A_Y(e)\widehat{\ell}_\theta(e)/\iota_e$, without an extra $1/B$ factor. This is an alternative sampling design, not the design analyzed in the main text.

B Proofs and Statistical Properties

B.1 Assumptions and conditional target preservation

Fix a context x and collection round t. The analysis uses a finite set of complete terminating trajectories, $\rho(y \mid x) > 0$, $\pi_t(y \mid x) > 0$, $0 \le \alpha < 1$, and $\tau > 0$. Candidate positions are independent draws from Eq. (6); the loss and collection weights are finite. Pair feedback has expectation $P_x(y \succ y')$, with reciprocity and $P_x(y \succ y) = 1/2$. Any active-pool allocation assigns positive mass to every pair with $A_Y(e) > 0$. During a round, all these collection quantities are fixed while π_θ changes.

Proof of Theorem 3.1. Conditionally on an active pool Y, one corrected draw has expectation

$$\sum_{e \in \mathcal{E}_Y} M_Y^*(e) \frac{A_Y(e)}{M_Y^*(e)} \ell_\theta(e) = \sum_{e \in \mathcal{E}_Y} A_Y(e)\ell_\theta(e) =: U_Y(\theta).$$

Averaging B independent draws has the same expectation. Set $U_Y = 0$ for a pool without valid pairs. Reciprocity gives $\ell_\theta(y, y') = \ell_\theta(y', y)$. Since every unordered pair of positions has the same

independent proposal law,

$$\mathbb{E}_Y U_Y(\theta) = \sum_{y \neq y'} S_{t,\lambda}(y) S_{t,\lambda}(y') \frac{g_t(y) g_t(y')}{S_{t,\lambda}(y) S_{t,\lambda}(y')} \ell_\theta(y, y') \tag{B.1}$$

$$= \sum_{y,y'} g_t(y) g_t(y') \ell_\theta(y, y') - \sum_y g_t(y)^2 \ell_\theta(y, y) \tag{B.2}$$

$$= Z_t(x)^2 \mathcal{L}_{t,x}(\theta) - \frac{1}{4\tau^2} \sum_y g_t(y)^2. \tag{B.3}$$

The positive multiplier and subtracted constant do not depend on the optimized policy parameters, so the context-wise minimizers coincide. □

For a context distribution μ, the expected training objective is

$$\mathbb{E}_{x\sim\mu}\big[Z_t(x)^2 \mathcal{L}_{t,x}(\theta)\big] - \mathbb{E}_{x\sim\mu} c_t(x), \qquad c_t(x) = \frac{1}{4\tau^2} \sum_y g_t(y \mid x)^2. \tag{B.4}$$

Thus its gradient, when differentiation is valid, is $\mathbb{E}_x[Z_t(x)^2 \nabla_\theta \mathcal{L}_{t,x}]$, not generally $\mathbb{E}_x \nabla_\theta \mathcal{L}_{t,x}$. The conditional correspondence becomes a simultaneous optimum only when the policy class can realize the conditional optima for all contexts.

B.2 Exact conditional optimizer and Nash correspondence

For a fixed positive opponent q, suppress the context and define $w_q(y) = \mathbb{E}_{y'\sim q} P(y \succ y')$ and $f(y) = \log(\pi(y)/\rho(y))$. Reciprocity implies $\mathbb{E}_q w_q = 1/2$. Expanding Eq. (A.1) under $q \times q$ yields

$$\mathcal{L}_q(\pi) = 2\,\mathrm{Var}_q(f) - \frac{4}{\tau}\,\mathrm{Cov}_q(f, w_q) + \frac{1}{4\tau^2} \tag{B.5}$$

$$= 2\,\mathrm{Var}_q\Big(f - \frac{w_q}{\tau}\Big) + \frac{1}{4\tau^2} - \frac{2}{\tau^2}\,\mathrm{Var}_q(w_q). \tag{B.6}$$

The minimum is achieved exactly when $f - w_q/\tau$ is constant on the trajectory space. Normalizing the resulting policy gives the unique conditional optimizer

$$\pi^q(y) = \frac{\rho(y) \exp(w_q(y)/\tau)}{\sum_u \rho(u) \exp(w_q(u)/\tau)}. \tag{B.7}$$

This is the online IPO optimizer associated with the preference-game connection in Calandriello et al. (2024). Theorem 3.1 shows that corrected pair allocation preserves this conditional map.

Proof of the Nash part of Corollary 3.2. At a fixed point, $\pi^* = \pi^{q^*}$ and $q^* \propto (\pi^*)^{1-\alpha}\rho^\alpha$. Equation (B.7) then gives

$$q^*(y) \propto \rho(y) \exp\left(\frac{1-\alpha}{\tau} w_{q^*}(y)\right) = \rho(y) \exp\left(\frac{w_{q^*}(y)}{\eta}\right).$$

The right side is the unique maximizer of $\mathbb{E}_{y\sim u} w_{q^*}(y) - \eta\,\mathrm{KL}(u\|\rho)$ over u. Hence q^* is a best response to itself in Eq. (1). Reciprocity gives $G^\eta(u, v) + G^\eta(v, u) = 1$, so it is also the minimizing player's best response. Therefore (q^*, q^*) is a Nash equilibrium. Conversely, any such q^* defines $\pi^* \propto \rho \exp(w_{q^*}/\tau)$ whose geometric transform is q^*, and Eq. (B.7) makes it a fixed point. □

For completeness, the regularized symmetric equilibrium exists and is unique in this finite setting. The continuous map $q \mapsto \rho \exp(Pq/\eta)/\sum_y \rho(y) \exp((Pq)_y/\eta)$ maps the compact simplex into itself, so it has a fixed point. For two fixed points q, r, subtracting their log fixed-point equations and taking the inner product with $q - r$ gives

$$\eta\big[\mathrm{KL}(q\|r) + \mathrm{KL}(r\|q)\big] = (q - r)^\top P (q - r) = 0.$$

The last equality uses $P + P^\top = \mathbf{1}\mathbf{1}^\top$ and $\mathbf{1}^\top(q - r) = 0$. Both divergences are nonnegative, and equality forces $q = r$. This also identifies a unique conditional policy fixed point through Eq. (B.7). Existence and uniqueness of a fixed point do not imply that iterating an arbitrary optimizer converges to it.

B.3 CONTINUATION-VALUE INTERPRETATION

A valid prefix is a union of complete-trajectory outcomes sharing the same initial tokens. Each step ends at an unambiguous boundary, so conditioning on $s \oplus z$ corresponds to the subset reached by that next step. Let

$$H_x(s) = \mathbb{E}_{u\sim\rho(\cdot|x,s)} \exp\big(w_x(s \oplus u)/\tau\big), \qquad V_x(s) = \tau \log H_x(s).$$

Summing Eq. (B.7) over completions of a prefix gives

$$\pi^*(s \mid x) = \rho(s \mid x)\frac{H_x(s)}{H_x(\varnothing)}.$$

Consequently,

$$\frac{\pi^*(z \mid x, s)}{\rho(z \mid x, s)} = \frac{H_x(s \oplus z)}{H_x(s)}, \qquad \beta \log \frac{\pi^*(z \mid x, s)}{\rho(z \mid x, s)} = \frac{\beta}{\tau}\big[V_x(s \oplus z) - V_x(s)\big], \tag{B.8}$$

which proves Eq. (15). Equivalently, the variational identity for the log moment-generating function gives

$$V_x(s) = \max_{\xi} \left\{\mathbb{E}_{u\sim\xi} w_x(s \oplus u) - \tau\, \mathrm{KL}\big(\xi \| \rho(\cdot \mid x, s)\big)\right\}. \tag{B.9}$$

Thus $V_x(s)$ is a reference-regularized expected terminal preference payoff. For candidates sharing the same prefix, $V_x(s)$ cancels in their reward difference. At a terminal trajectory y, $V_x(y) = w_x(y)$; summing reward increments through answer and termination recovers $\beta[w_x(y) - V_x(\varnothing)]/\tau$. The interpretation concerns the specified preference payoff, rather than a probability of local factual correctness.

B.4 OPTIMALITY OF THE PAIR ALLOCATION

Proof of Proposition 3.3. Write $Q = \bar{Q}_Y$, $P = P_Y$, $L = (1-\gamma)P$, and $U = P + \gamma Q$. On the valid support, $Q, P > 0$. The anchor P is feasible. For $0 < \gamma < 1$, minimizing $\sum_e Q_e^2/M_e$ is a strictly convex optimization problem on the feasible simplex. The stationarity condition for an interior coordinate is $-Q_e^2/M_e^2 + \zeta = 0$, giving $M_e = Q_e/\sqrt{\zeta}$. With the box constraints, the Karush–Kuhn–Tucker conditions yield $M_e^* = \mathrm{clip}(Q_e/c, L_e, U_e)$ with $c = \sqrt{\zeta}$ chosen to sum to one. This is the unique minimizer of $\mathfrak{J}_Y$.

To prove minimax optimality, first suppose some coordinate is clipped at its upper bound with $Q_e/c \geq U_e$. Then the maximum ratio at M^* is $\max_e Q_e/U_e$ over upper-clipped coordinates, which is at least c. All other ratios are at most c. Every feasible allocation has ratio at least Q_e/U_e at each such coordinate, so none has a smaller maximum.

Otherwise, no coordinate is clipped from above and $M_e^* = \max\{L_e, Q_e/c\}$. Since $\sum_e L_e = 1-\gamma < 1$, at least one coordinate exceeds its lower bound and the maximum ratio equals c. Any allocation with all ratios strictly below c must exceed Q_e/c on those coordinates while remaining at least L_e elsewhere. Its total mass would exceed one. Thus no feasible allocation has a smaller maximum in this case either. The endpoint $\gamma = 0$ has only the feasible distribution P; at $\gamma = 1$, Q is feasible and attains the universal lower bounds $\mathfrak{C} = \mathfrak{J} = 1$. □

For $0 \leq \gamma < 1$, exact target allocation Q is feasible precisely when

$$1 - \gamma \leq Q_e/P_e \leq (1-\gamma)^{-1} \quad \text{for every valid pair } e. \tag{B.10}$$

For $\gamma > 0$, the linear allocation $(1-\gamma)P + \gamma Q$ is feasible and has worst ratio at most $1/\gamma$. Optimality therefore implies $\mathfrak{C}_Y(M_Y^*) \leq 1/\gamma$.

B.5 PROPOSAL AND PAIR-ESTIMATION SECOND MOMENTS

Finite-pool variation. For $q = g/Z$ and proposal S, define

$$\Gamma(q, S) = \sum_y \frac{q(y)^2}{S(y)} = 1 + \chi^2(q \| S). \tag{B.11}$$

If $|\ell_\theta(y, y')| \le L_{\max}$, the symmetric degree-two kernel $H(y, y') = [g(y)/S(y)][g(y')/S(y')]\mathbf{1}\{y \neq y'\}\ell_\theta(y, y')$ has second moment at most $L_{\max}^2 Z^4 \Gamma(q, S)^2$. The standard degree-two U-statistic variance bound (Hoeffding, 1992) gives the conservative bound

$$\mathrm{Var}_Y\, U_Y(\theta) \le \frac{4L_{\max}^2 Z^4}{K}\Gamma(q, S)^2. \tag{B.12}$$

This quantifies the effect of the chosen candidate proposal; it does not make λ an optimizer of Γ.

Pair and feedback variation. Conditionally on an active pool, suppose $\mathbb{E}[\widehat{\ell}_e^2 \mid Y, e] \le H_Y^2$, with independent feedback observations across pair draws. Then

$$\mathrm{Var}(\widehat{U}_Y \mid Y) \le \frac{\kappa_Y^2 H_Y^2}{B}\mathfrak{J}_Y(M). \tag{B.13}$$

For deterministic fixed per-pair losses, the exact variance is

$$\mathrm{Var}(\widehat{U}_Y \mid Y) = \frac{1}{B}\left[\sum_e \frac{A_Y(e)^2\ell_e^2}{M(e)} - U_Y^2\right]. \tag{B.14}$$

Thus minimizing $\mathfrak{J}_Y$ minimizes the common bound in Eq. (B.13), not necessarily the exact loss variance in Eq. (B.14). The latter depends on ℓ_e^2 as well as pair masses. The total variance separates into finite-pool variation and expected conditional pair/feedback variation by the law of total variance.

Anchor-relative gradient second moment. Let $v_e \ge 0$ be the conditional second moment of the gradient of the unweighted pair loss. For one corrected draw, write $S_v(M) = \sum_e A_Y(e)^2 v_e/M(e)$. For nonzero v,

$$\sup_{v\ge 0,\, v\neq 0} \frac{S_v(M)}{S_v(P_Y)} = \max_e \frac{P_Y(e)}{M(e)} \le \frac{1}{1-\gamma}, \qquad 0 \le \gamma < 1. \tag{B.15}$$

The equality follows because the ratio is a weighted average of $P_Y(e)/M(e)$ and its maximum is attained by concentrating v on a maximizing coordinate. The inequality follows from the anchor floor. This is a worst-case second-moment factor, not a reduction in centered gradient variance. The actual gradient-second-moment optimum would additionally depend on the unknown v_e.

C QA Experimental Protocol

Data and initialization. The reported evaluation uses the full official WebQSP and CWQ test sets (Yih et al., 2016; Talmor & Berant, 2018), based on Freebase (Bollacker et al., 2008). DPO, IPO, IPO-MD, and CO-OIPO are initialized directly from the same Qwen3-8B base checkpoint (Yang et al., 2025), with a frozen copy as reference. They are trained separately on each dataset and share the training questions within that dataset.

Optimization and inference. The reported QA optimizer is AdamW (Loshchilov & Hutter, 2019), with learning rate 10^{-6} and effective batch size 16 contexts: one context per GPU, four GPUs, and four gradient-accumulation steps. CO-OIPO uses three online rounds with $\alpha = 0.5$, $\tau = 0.1$, $\gamma = 0.6$, proposal mixing $\lambda = 0.5$, $K = 4$ candidates per pool, and $B = 2$ pair draws per active pool. At inference, a Qwen3-8B instruction model proposes eight next-step candidates under the selected prefix, with at most four steps. The final-checkpoint implicit PRM uses Eq. (5) directly.

Preference feedback. Each sampled QA training pair is sent to DeepSeek-V4-Flash in one presentation order. Left win, tie, and right win are encoded as 1, $1/2$, and 0, respectively. Failed or invalid calls are retried without deleting the pair. The reciprocal-kernel condition used by the analysis is specified in Appendix A; the QA training pairs use the single presentation order described above.

Metrics and baselines. Hit measures whether the predicted answer set intersects the gold answer set. F1 measures their precision–recall agreement. Table 1 includes Qwen3 and Llama 3 models (Yang et al., 2025; Grattafiori et al., 2024); SubgraphRAG, DCTR, and ParallaxRAG (Li et al., 2025; Cao et al., 2026; Liu et al., 2026a); Skywork-Reward-V2, DeepSeek-GRM, and RM-R1 (Liu et al.,

2026c; 2025; Chen et al., 2026); and ThinkPRM, VersaPRM, PRInTS, and DPRM (Khalifa et al., 2026; Zeng et al., 2025; Lee et al., 2026; Wang et al., 2026). Published results are used where available, with additional evaluations for other entries. The four preference-learning rows share the stated initialization and dataset-specific training questions.

D Frozen-Pool Sampling Diagnostics

D.1 Evaluation units and fixed quantities

This diagnostic uses the same 4,750 active WebQSP candidate pools as Table 4, with $K = 4$ candidates per pool and $B = 2$ pair draws. Candidate pools, model checkpoints, unnormalized pair masses, and locally computed per-pair preference labels are held fixed when comparing allocations. These diagnostic labels are not the DeepSeek-V4-Flash labels used in QA training.

For each pool, Anchor uses P_Y, Linear uses $(P_Y + \bar{Q}_Y)/2$, and CO-OIPO uses M_Y^*, with the common box parameter $\gamma = 0.5$. All three satisfy the probability constraints on every evaluated pool, use the full correction A_Y/M_Y, and share the same deterministic pair losses. Both the coverage statistics and loss-estimation errors are computed over this same set of 4,750 active pools.

D.2 Exact MSE and aggregate results

The diagnostic evaluates Eq. (B.14) with $B = 2$ by summing over every valid pair. Thus, it computes the exact conditional MSE of the corrected loss estimator for the frozen losses, rather than estimating the MSE from a finite number of random pair trials. It does not include additional stochastic judge variation.

Table 6: **Exact corrected-loss estimation error on the 4,750 frozen WebQSP candidate pools.** The same evaluation population is used for the coverage statistics reported in Table 4. Lower is better for both metrics.

Sampler	Mean MSE ↓	Median MSE ↓
Anchor	0.439862	0.015215
Linear	0.250842	0.012568
CO-OIPO	**0.224850**	**0.011748**

Across the same 4,750 pools, mean $\mathfrak{C}_Y - 1$ is 0.3297, 0.1134, and 0.0084 for Anchor, Linear, and CO-OIPO, respectively, while mean $\mathfrak{J}_Y - 1$ is 0.0572, 0.0118, and 0.0008. CO-OIPO reduces mean MSE by 48.9% relative to Anchor and by 10.4% relative to Linear.

The aggregate MSE improvement does not hold on every pool: CO-OIPO has higher loss MSE on 1,006 pools relative to Anchor and on 1,240 pools relative to Linear. Numerical equality is assessed with $|u - v| \le 10^{-12} + 10^{-9}\max(|u|, |v|)$. These observations are compatible with Proposition 3.3, which optimizes pair-distribution criteria rather than the loss-dependent quantity in Eq. (B.14).

E Controlled Four-Action Preference Learning

E.1 Environment, initialization, and optimization

In action order A, B, C, N, the row action is preferred to the column action with probabilities

$$P_{\mathrm{sym}} = \begin{pmatrix} 0.5 & 1 & 0 & 0.9 \\ 0 & 0.5 & 1 & 0.9 \\ 1 & 0 & 0.5 & 0.9 \\ 0.1 & 0.1 & 0.1 & 0.5 \end{pmatrix}, \qquad P_{\mathrm{asym}} = \begin{pmatrix} 0.5 & 0.8 & 0.4 & 0.9 \\ 0.2 & 0.5 & 0.7 & 0.9 \\ 0.6 & 0.3 & 0.5 & 0.9 \\ 0.1 & 0.1 & 0.1 & 0.5 \end{pmatrix}. \tag{E.1}$$

The prompt is "Select one of four actions. Reply with exactly one symbol: A, B, C, or N." It reveals neither the preference matrix nor the role of N. Each queried pair receives hard feedback sampled from $\mathrm{Bernoulli}(P_{ij})$.

The frozen Qwen3-8B base is adapted with rank-16 LoRA, scaling 32, and zero dropout on the query, key, value, output, gate, up, and down projections. Each complete response is an action token followed by EOS. Two fixed first-token logit biases calibrate the reference to $\rho = (0.3, 0.3, 0.3, 0.1)$ and the initial zero-adapter policy to $\pi_0 = (0.225, 0.45, 0.225, 0.1)$. Action probabilities are obtained by normalizing the probabilities of these four calibrated complete responses. This defines the four-action policy. Training retains both calibration biases in $f_i = \log p_{\theta,\mathrm{cal}}(y_i \mid x) - \log p_{\mathrm{ref,cal}}(y_i \mid x)$; the difference $f_i - f_j$ equals $\log(\pi_i/\rho_i) - \log(\pi_j/\rho_j)$.

Offline DPO uses 1,000 comparisons generated before training, independently of the reference and training policies. With probability 0.8, it draws uniformly from AN, BN, CN; otherwise it draws uniformly from AB, BC, CA. Each pair receives one Bernoulli label. A single shuffled pass uses batch size one and one update per record, with loss $-\log \sigma[\beta_{\mathrm{DPO}}(2v-1)(f_i - f_j)]$ and $\beta_{\mathrm{DPO}} = 0.5$.

CO-OIPO uses $\alpha = 0.5$, $\tau = 0.1$, $\lambda = 0.95$, $\gamma = 0.8$, $K = 8$, and $B = 1$. Each informative update collects a pool from $S = 0.05\pi + 0.95\rho$, retains duplicate indices, constructs the allocation from its pair masses, and queries one hard label. Its corrected loss is

$$\frac{A_Y(i,j)}{M_Y^*(i,j)} \left[f_i - f_j - \frac{2v-1}{2\tau} \right]^2 .$$

Pair allocation does not access the true preference probabilities. Identical-action pairs are omitted. The development runs had no zero-contribution pools and used 8,000 candidate draws for 1,000 labels. Candidate counts for the additional seeds are not tabulated in the supplied three-seed summary.

Both methods use AdamW with betas $(0.9, 0.999)$, zero weight decay, and no gradient clipping. The selected peak learning rate is 2×10^{-6}, with 5% warmup and cosine decay to 2×10^{-7}. Training ends at 1,000 labels and 1,000 updates. Learning rates 5×10^{-7} and 2×10^{-6} were checked for both methods during development; CO-OIPO received a broader sampling-parameter search on the symmetric game, and those sampling parameters were transferred to the asymmetric game. Seed 123 is the development seed; 42 and 43 are additional replications. All results in Table 3 use the fixed final checkpoint rather than a selected intermediate checkpoint.

E.2 Evaluation and equilibrium targets

The output policy is evaluated by $E(\pi) = \max_a (P\pi)_a - 1/2$. Both methods use the same transformed distribution $q \propto \pi^{0.5}\rho^{0.5}$ and $\eta = 0.2$ for their regularized Nash gaps. Reciprocity implies $\min_v G^\eta(q, v) = 1 - \max_u G^\eta(u, q)$, so

$$\mathrm{Gap}_\eta(q) = 2 \left[\eta \log \sum_a \rho_a \exp((Pq)_a/\eta) + \eta\, \mathrm{KL}(q\|\rho) \right] - 1. \tag{E.2}$$

This is the gap of the regularized game, whereas $E(\pi)$ concerns the unregularized exploitability of the output policy.

The regularized equilibria are approximately

$$q^*_{\mathrm{sym}} = (0.328395, 0.328395, 0.328395, 0.014814),$$
$$q^*_{\mathrm{asym}} = (0.414525, 0.248489, 0.321856, 0.015131).$$

The corresponding exact output-policy fixed points are obtained from $\pi^* \propto (q^*/\rho^\alpha)^{1/(1-\alpha)}$:

$$\pi^*_{\mathrm{sym}} = (0.332656, 0.332656, 0.332656, 0.002031),$$
$$\pi^*_{\mathrm{asym}} = (0.508594, 0.182760, 0.306613, 0.002033).$$

The unregularized equilibria are $(1/3, 1/3, 1/3, 0)$ and $(1/3, 1/6, 1/2, 0)$, respectively. These distributions are different evaluation targets; in particular, vanishing regularized gap does not require zero output-policy exploitability.

E.3 All three final endpoints

Tables 7 and 8 reproduce all final endpoints in the three-seed summary. Action probabilities are rounded percentages and may not sum to exactly 100 after display rounding. The metrics use the

Table 7: Symmetric game: fixed final checkpoint after 1,000 feedbacks and updates. Action probabilities are percentages; E and gap are unscaled.

Seed	Method	A	B	C	N	$E(\pi)$	Gap
123	DPO	38.85	39.80	21.30	0.048	0.09266	0.01232
123	CO-OIPO	33.90	33.50	32.35	0.247	0.00672	0.00009
42	DPO	39.24	36.22	24.34	0.208	0.06020	0.00600
42	CO-OIPO	31.80	22.30	45.81	0.090	0.07040	0.01279
43	DPO	37.86	34.67	27.30	0.173	0.03752	0.00274
43	CO-OIPO	32.33	37.82	29.60	0.254	0.04211	0.00160

Table 8: Asymmetric game: fixed final checkpoint after 1,000 feedbacks and updates, with the same reporting convention as Table 7.

Seed	Method	A	B	C	N	$E(\pi)$	Gap
123	DPO	52.09	33.59	14.27	0.052	0.08670	0.01819
123	CO-OIPO	55.26	20.79	23.85	0.090	0.03888	0.00209
42	DPO	47.52	27.83	24.40	0.252	0.06011	0.00443
42	CO-OIPO	62.68	14.44	22.72	0.169	0.03448	0.00373
43	DPO	38.81	40.17	20.89	0.124	0.10012	0.01798
43	CO-OIPO	46.01	28.52	25.25	0.217	0.06117	0.00468

unrounded policies in the experiment records.Table 3 reports arithmetic means across the three seeds; Tables 7 and 8 provide the individual endpoints.

In the symmetric game, CO-OIPO improves both metrics at seed 123, has a smaller gap but larger exploitability at seed 43, and has larger values on both at seed 42. In the asymmetric game it improves both metrics in all three seeds. The means summarize these runs; they are not significance tests. Equal numbers of labels and updates also do not equalize feedback information, candidate computation, or development effort between offline DPO and online CO-OIPO.

E.4 Development checkpoint trajectories

The following tables retain the original seed-123 trajectories to show development-run behavior. They are not additional independent runs and are not used for final-checkpoint selection. The finer displayed precision comes from the original development tables, whereas Tables 7 and 8 follow the precision of the new three-seed summary.

Table 9: Symmetric four-action game: development-run action probabilities (%), output exploitability, and common regularized Nash gap at seed 123. Metrics are unscaled; $\alpha = 0.5$ and $\eta = 0.2$.

Method	Updates	A	B	C	N	$E(\pi) \downarrow$	$\mathrm{Gap}_\eta(q) \downarrow$
DPO	0	22.500	45.000	22.500	10.000	0.152500	0.059693
DPO	200	25.447	49.749	24.195	0.610	0.130212	0.018555
DPO	400	36.452	39.119	24.370	0.059	0.073984	0.006921
DPO	600	34.508	40.840	24.586	0.065	0.081533	0.007201
DPO	800	36.045	42.748	21.104	0.104	0.108635	0.012975
DPO	1000	38.850	39.798	21.304	0.048	0.092660	0.012323
CO-OIPO	0	22.500	45.000	22.500	10.000	0.152500	0.059693
CO-OIPO	200	48.323	26.631	24.376	0.669	0.111136	0.015948
CO-OIPO	400	34.556	11.291	54.052	0.100	0.116725	0.050675
CO-OIPO	600	24.561	32.152	43.152	0.134	0.093492	0.008211
CO-OIPO	800	29.330	33.363	37.139	0.168	0.039714	0.001420
CO-OIPO	1000	33.903	33.498	32.351	0.247	0.006722	0.000088

Table 10: Asymmetric four-action game: development-run action probabilities (%), output exploitability, and common regularized Nash gap at seed 123. Metrics are unscaled; $\alpha = 0.5$ and $\eta = 0.2$.

Method	Updates	A	B	C	N	$E(\pi) \downarrow$	$\mathrm{Gap}_\eta(q) \downarrow$
DPO	0	22.500	45.000	22.500	10.000	0.152500	0.074832
DPO	200	35.861	43.881	19.469	0.789	0.115331	0.025280
DPO	400	45.948	39.314	14.644	0.094	0.103676	0.022444
DPO	600	46.546	37.695	15.705	0.055	0.097598	0.019911
DPO	800	50.060	35.835	14.048	0.057	0.093682	0.020394
DPO	1000	52.088	33.587	14.272	0.052	0.086698	0.018192
CO-OIPO	0	22.500	45.000	22.500	10.000	0.152500	0.074832
CO-OIPO	200	52.087	24.533	22.820	0.560	0.053018	0.004208
CO-OIPO	400	55.908	11.324	32.484	0.284	0.034395	0.003512
CO-OIPO	600	53.731	14.269	31.864	0.137	0.025740	0.001094
CO-OIPO	800	48.606	19.145	32.164	0.085	0.025613	0.000567
CO-OIPO	1000	55.264	20.793	23.853	0.090	0.038883	0.002089

E.5 Separate proposal-mixing and pair-allocation ablation

The ablation uses the asymmetric game, the same calibrated Qwen3-8B backend, rank-16 LoRA, and the AdamW schedule above. Its initialization differs from the DPO comparison: $\pi_0 = (0.075, 0.75, 0.075, 0.1)$ with reference $\rho = (0.3, 0.3, 0.3, 0.1)$. All variants use $\alpha = 0.5$, $\tau = 0.1$, $K = 8$, and $B = 1$, and are evaluated with $\eta = 0.2$. Proposal mixing switches λ between 0 and 0.75; pair allocation switches γ between 0 and 0.8. Every update retains the full A_Y/M_Y correction.

Training collects until 1,000 informative labels and updates, recording the candidate cost of zero-contribution pools separately. At a fixed round in this single-context game, conditioning on an active pool divides the expectation in Theorem 3.1 by $\Pr(\kappa_Y > 0)$. This positive, collection-dependent factor preserves the conditional minimizers but changes the expected update scale. The table below therefore records both the informative budget and total candidate cost; the unconditional estimator in Eq. (4) instead retains zero contributions in its original denominator.

The ablation uses seeds 42, 123, and 2024 with independent candidate, pair, and feedback streams. Table 2 reports means; Table 11 gives their sample standard deviations; Table 12 gives individual endpoints and candidate costs. The joint setting has the smallest gap in each of the three seeds, although its exploitability is not smallest in every seed.

Table 11: Asymmetric-game ablation over seeds 42, 123, and 2024: mean $\pm$ sample standard deviation.

Proposal mixing	Pair allocation	$E(\pi) \downarrow$	Gap $\downarrow$
Off	Off	0.032820 ± 0.005336	0.003077 ± 0.001822
On	Off	0.036656 ± 0.017415	0.002920 ± 0.001208
Off	On	0.030022 ± 0.011123	0.003253 ± 0.001754
On	On	0.027967 ± 0.005850	0.000748 ± 0.000482

Table 12: Individual asymmetric-game ablation endpoints. Every run uses 1,000 labels and updates. Candidate draws include all zero-contribution collection attempts. Switches indicate proposal mixing and optimized pair allocation.

Seed	Mixing	Allocation	$E(\pi)$	Gap	Candidates	Zero pools
42	0	0	0.027116	0.002665	8,512	64
42	1	0	0.033453	0.003611	8,000	0
42	0	1	0.030670	0.001228	8,144	18
42	1	1	0.023975	0.000363	8,000	0
123	0	0	0.033653	0.001496	8,240	30
123	1	0	0.055450	0.003624	8,000	0
123	0	1	0.040806	0.004256	8,368	46
123	1	1	0.034682	0.000592	8,000	0
2024	0	0	0.037689	0.005070	8,216	27
2024	1	0	0.021065	0.001524	8,008	1
2024	0	1	0.018589	0.004276	8,264	33
2024	1	1	0.025242	0.001289	8,008	1