ORDIPO: ORTHOGONAL DISTILLATION FOR PREFERENCE OPTIMIZATION IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowledge distillation provides token-level teacher guidance that can complement pairwise preference learning for language model alignment. However, closer imitation of the teacher's predictive distributions does not necessarily improve alignment with the target preferences. Through controlled gradient-component interventions, we show that the distillation residual orthogonal to an estimated preference-gradient direction retains useful generative guidance. Motivated by this observation, we propose Orthogonal Distillation for Preference Optimization (OrDiPO), which treats Direct Preference Optimization as the primary objective and chosen-prefix distillation as an auxiliary signal. Specifically, OrDiPO uses an exponential moving average of past preference gradients to define the projection normal. It removes the distillation-gradient component parallel to this normal and combines the remaining orthogonal residual with the unmodified current preference gradient. Experiments on UltraMedical-Preference and SHP–AskDocs demonstrate improvements over DPO and DPO+KD in both reference-corrected pair accuracy and generation win rate on internal test sets, with further evidence from external benchmarks. Additional experiments on HelpSteer3-Preference support applicability in a general-domain setting, while ablations and gradient analyses support the roles of orthogonal guidance and preference-gradient averaging.

1 INTRODUCTION

Learning from response preferences is an established paradigm for adapting language models to desired behaviors (Ouyang et al., 2022). In this setting, Direct Preference Optimization (DPO) learns a policy from pairwise response comparisons (Rafailov et al., 2023), and knowledge distillation can complement this supervision with fine-grained predictive distributions from a teacher model (Hinton et al., 2015; Agarwal et al., 2024). Although combining these signals holds promise for enhancing response generation, their supervisory roles differ: preference labels specify relative rankings, whereas distillation encourages matching the teacher's predictive distributions. Closer imitation therefore need not translate to better satisfaction of the underlying preference criteria. A fundamental challenge thus lies in exploiting the teacher's generative guidance while preserving the primacy of preference optimization.

Existing approaches frequently combine preference learning with supervised generation (Liu et al., 2024; Hong et al., 2024). Teacher-guided approaches transfer aligned predictive distributions or teacher-derived advantages (Gao et al., 2025), incorporate teacher and reference distributions into preference formulations (Zhang et al., 2025), or use teacher values for reward shaping (Kwon et al., 2026). In parallel, gradient surgery addresses conflicts between task updates (Yu et al., 2020), including joint supervised–preference training for knowledge editing (Nguyen et al., 2026), while auxiliary-learning methods regulate secondary gradients according to a designated primary objective (Dery et al., 2021; Hsieh et al., 2024). Building on these advances, we ask whether the generative benefits of chosen-prefix distillation can be retained when its gradient is restricted to the subspace orthogonal to an estimated preference-gradient direction.

Our controlled preliminary experiments show that closer teacher imitation on chosen-response prefixes can improve generation quality while reducing preference accuracy. Because joint loss weighting scales the entire distillation gradient, we examine its directional contributions. Gradient-

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.

component interventions show that the orthogonal residual improves both preference accuracy and response quality over full-gradient guidance in this setting. Comparisons with norm-matched full gradients and random orthogonal controls further suggest that these gains are not explained solely by auxiliary-gradient magnitude or random orthogonal perturbations.

Motivated by these observations, we introduce **Orthogonal Distillation for Preference Optimization (OrDiPO)**, which treats DPO as the primary objective and chosen-prefix distillation as auxiliary guidance. Inspired by primary–auxiliary optimization (Hsieh et al., 2024), OrDiPO integrates teacher guidance at the gradient level rather than through a scalar mixture of training losses. It combines the current preference gradient with the component of the distillation gradient orthogonal to a historical preference-gradient estimate, retaining teacher-derived guidance within this complementary subspace.

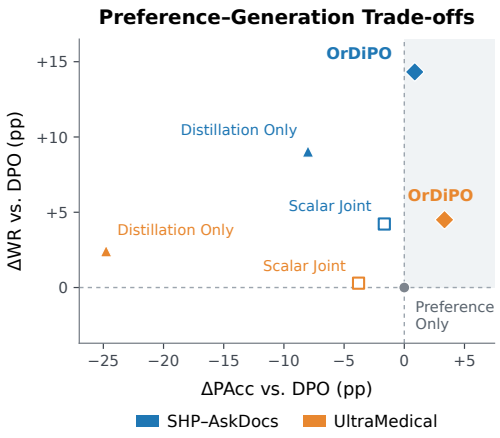

Figure 1: **Preference–generation trade-offs.** Changes relative to each dataset's DPO baseline. OrDiPO improves both preference accuracy and generation win rate.

Our main experiments span UltraMedical-Preference (Zhang et al., 2024) and SHP–AskDocs (Ethayarajh et al., 2022) across two model families, assessing reference-corrected pair accuracy (PAcc) and generation win rate (WR) on internal test sets. Figure 1 summarizes the Qwen3-8B-Base results. External evaluations use K-QA (Manes et al., 2024) and HealthBench (Arora et al., 2025). Independent training on HelpSteer3-Preference (Wang et al., 2026) examines applicability beyond health (Appendix J), while ablations assess distillation, orthogonal projection, and preference-gradient averaging.

Our contributions are threefold:

- We empirically characterize preference–imitation trade-offs in a controlled setting and demonstrate the utility of orthogonal distillation residuals through gradient-component interventions.
- We propose OrDiPO, a preference-primary alignment framework that pairs DPO with cached chosen-prefix teacher distributions and integrates teacher guidance through an orthogonal auxiliary update.
- We demonstrate improved preference–generation trade-offs on two health-related datasets, supported by external evaluations and component ablations, and provide evidence of applicability beyond health through independent general-domain training.

2 Preliminary

In this section, we empirically examine the interplay between teacher guidance and preference learning. We first characterize preference–imitation trade-offs (Section 2.1), then investigate the utility of the distillation residual orthogonal to an estimated preference-gradient direction (Section 2.2).

2.1 Preference Learning and Teacher Imitation

Knowledge distillation can complement preference learning with teacher predictive distributions (Hinton et al., 2015; Agarwal et al., 2024). While pairwise preference objectives supervise the relative ranking between chosen responses y^+ and rejected responses y^-, our distillation objective matches token-level teacher distributions conditioned on chosen-response prefixes $(x, y^+_{<\tau})$. These distributions provide informative generative priors, yet conditioning solely on chosen prefixes does not ensure that matching them reinforces the target preference margin. This raises a practical question: *when teacher guidance improves generation, does closer imitation also improve preference learning?*

We examine this question using Qwen3-8B-Base on UltraMedical-Preference (Zhang et al., 2024). We compare Direct Preference Optimization (DPO) alone against scalar-weighted chosen-prefix knowledge distillation (DPO+KD), optimizing $\mathcal{L}_P(\theta) + \lambda\mathcal{L}_D(\theta)$. To distinguish teacher-distribution guidance from chosen-token supervision, we also evaluate DPO+SFT, optimizing $\mathcal{L}_P(\theta)+\mu\mathcal{L}_{\mathrm{SFT}}(\theta)$, where $\mathcal{L}_{\mathrm{SFT}}$ is the cross-entropy loss on training chosen responses. We independently vary λ and μ over $\{0.03, 0.10, 0.30, 0.50\}$, with common initialization and data splits across joint-training configurations. DPO provides the unaugmented baseline, and sequential SFT→DPO provides a supervised warm-up comparison. We track evaluation distillation loss and internal-test PAcc, and report win, tie, and loss percentages against the corresponding test-set chosen responses. Further experimental and evaluation details are provided in Appendix D.

Figure 2 (right) shows that DPO+KD at $\lambda = 0.30$ achieves a higher win percentage and lower loss percentage than DPO and the evaluated SFT-based configurations, suggesting benefits beyond chosen-token supervision in this setting. These generation gains nevertheless coexist with lower PAcc. As λ increases, the evaluation distillation loss decreases (left), while PAcc falls overall from approximately 73% at $\lambda = 0.03$ to around 63% at $\lambda \in \{0.30, 0.50\}$ (middle). DPO+SFT retains higher PAcc at these larger weights but does not match the win percentages of DPO+KD at $\lambda \in \{0.10, 0.30\}$. Moreover, increasing λ from 0.30 to 0.50 further reduces the distillation loss while decreasing the win percentage. Closer imitation therefore does not uniformly improve preference agreement or generation outcomes, motivating the examination of gradient components in Section 2.2.

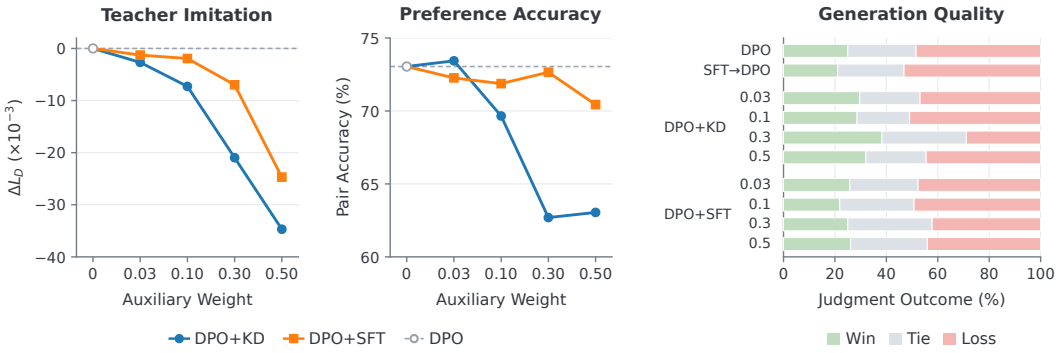

Figure 2: Teacher imitation and preference learning on UltraMedical-Preference. **Left**: change in held-out distillation loss. **Middle**: reference-corrected pair accuracy. **Right**: win, tie, and loss percentages against the corresponding held-out chosen responses.

2.2 Gradient Structure of Teacher Guidance

While scalar weighting modulates the overall strength of teacher imitation, it scales the entire distillation gradient without distinguishing its directional components. We therefore examine whether useful teacher guidance remains after removing its component along an estimated preference-gradient direction. For this diagnostic, we use an exponential moving average (EMA) of preceding preference gradients as the projection normal n_t and decompose the instantaneous distillation gradient $g_{D,t}$ into its parallel component and orthogonal residual $r_{D,t}$. The projection construction is specified in Section 3.2.

Table 1: Gradient-component interventions on UltraMedical-Preference with Qwen3-8B-Base.

Auxiliary Guidance	**PAcc (%)** ↑	**WR (%)** ↑
None	73.85	31.62
Full	72.70	35.14
Parallel	74.65	36.69
Norm-matched full	73.04	37.85
Random orthogonal	72.24	36.92
Orthogonal residual	**77.42**	**41.42**

Using the same student policy and dataset, we conduct controlled continuations from a shared intermediate DPO checkpoint. All branches start from identical model parameters, optimizer states, and EMA histories and follow the same optimizer-batch sequence.

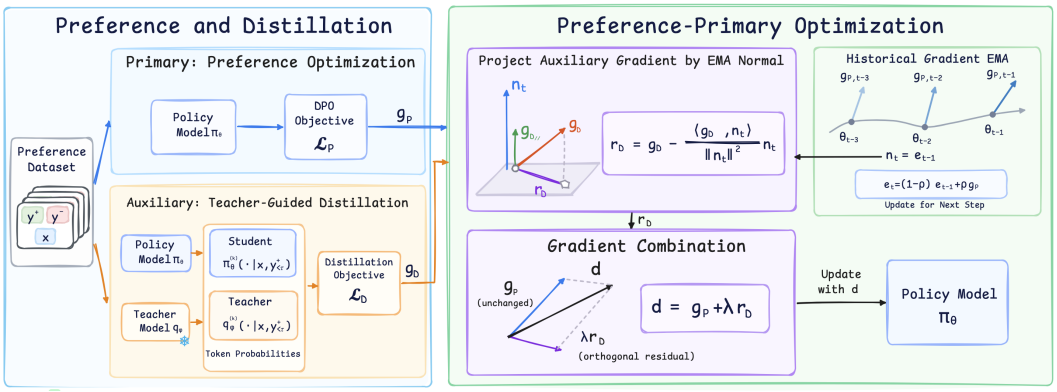

Figure 3: Overview of the OrDiPO pipeline. OrDiPO first pairs primary DPO with auxiliary chosen-prefix distillation from a frozen teacher. Next, it removes the component of the distillation gradient g_D parallel to the EMA of past preference gradients $n_t = e_{t-1}$. Finally, the resulting orthogonal residual r_D is combined with the unmodified current preference gradient g_P to form $d = g_P + \lambda r_D$ for the policy update, then the EMA incorporates g_P for the next iteration. Current-step subscripts are omitted from g_P, g_D, r_D, and d for clarity.

In each branch, the current preference gradient $g_{P,t}$ enters the combined direction unchanged, while the auxiliary contribution is omitted, retained in full, or restricted to the parallel component or orthogonal residual. To assess the influence of auxiliary-gradient magnitude and direction, we further evaluate two control baselines: a norm-matched full distillation gradient and a random direction orthogonal to n_t, both calibrated to match $\|r_{D,t}\|$ prior to optimizer preconditioning. Intervention details and geometric qualifications are provided in Appendices D and B.

Table 1 shows that the orthogonal residual achieves the highest PAcc and WR among the evaluated interventions, outperforming both preference-only continuation and full distillation. Although parallel-only guidance also improves both metrics over preference-only continuation, it remains below the residual. Norm-matched full and random orthogonal guidance improve WR but reduce PAcc relative to preference-only continuation. These comparisons support the residual's empirical utility in this controlled setting, without establishing a general guarantee of preference preservation.

3 Methodology

In this section, we introduce Orthogonal Distillation for Preference Optimization (OrDiPO), which treats preference optimization as the primary objective and teacher guidance as an auxiliary signal, as illustrated in Figure 3. We first define the training objectives in Section 3.1, then describe their preference-primary optimization in Section 3.2.

3.1 Preference and Distillation Objectives

Let $\mathcal{D} = \{(x_i, y_i^+, y_i^-)\}_{i=1}^N$ denote the training preference dataset, where y_i^+ is preferred to y_i^- under context x_i. We designate the trainable student policy as π_θ, a frozen reference policy initialized from the student prior to training as π_{ref}, and a fixed teacher model as q_ϕ. Only the trainable student parameters θ are updated.

Preference Objective. We adopt Direct Preference Optimization (DPO) (Rafailov et al., 2023) as the primary objective. Defining the reference-corrected sequence score as $s_\theta(x, y) = \log \pi_\theta(y \mid x) - \log \pi_{\text{ref}}(y \mid x)$, the preference loss is

$$\mathcal{L}_P(\theta) = -\mathbb{E}_{(x,y^+,y^-)\sim\mathcal{D}} \left[\log \sigma\big(\beta \left[s_\theta(x, y^+) - s_\theta(x, y^-)\right]\big)\right], \tag{1}$$

where σ denotes the logistic sigmoid function and $\beta > 0$ scales the preference margin. Sequence log-probabilities are computed by summing token log-probabilities over valid response positions without length normalization.

Teacher-Guided Distillation. The teacher provides next-token predictive distributions conditioned on chosen-response prefixes $y^+_{<\tau}$ concatenated with prompt x, assuming an aligned vocabulary and identical tokenization with the student. To reduce teacher-target storage, we retain the probabilities of the teacher's top-k tokens at each position with $k = 64$ and aggregate the residual probability mass into a single class OTHER. The student distribution follows the identical grouping. Both grouped distributions $q^{(k)}_\phi$ and $\pi^{(k)}_\theta$ are computed directly from full-vocabulary probabilities without renormalizing the retained tokens. The auxiliary distillation objective is formulated as

$$\mathcal{L}_D(\theta) = \mathbb{E}_{(x,y^+,y^-)\sim\mathcal{D}}\left[\frac{1}{T^+}\sum_{\tau=1}^{T^+} D_{\mathrm{KL}}\Big(q^{(k)}_\phi(\cdot \mid x, y^+_{<\tau}) \,\|\, \pi^{(k)}_\theta(\cdot \mid x, y^+_{<\tau})\Big)\right], \tag{2}$$

where T^+ counts valid chosen-response tokens. The loss is averaged within each chosen response and then over preference pairs. Teacher targets are precomputed and fixed, while student probabilities are recomputed at each training step. Distillation supervision is applied exclusively to the chosen response branch. Details of probability grouping, token masking, loss reduction, and the corresponding logit-gradient derivation are provided in Appendix A.

3.2 PREFERENCE-PRIMARY OPTIMIZATION VIA ORTHOGONAL PROJECTION

As shown in Figure 3 (right), OrDiPO combines the objectives at the gradient level, retaining the preference gradient and incorporating teacher guidance through an orthogonal distillation residual.

3.2.1 ORTHOGONAL DISTILLATION RESIDUAL

At iteration t, we compute $g_{P,t} = \nabla_\theta \widehat{\mathcal{L}}_{P,t}(\theta_t)$ and $g_{D,t} = \nabla_\theta \widehat{\mathcal{L}}_{D,t}(\theta_t)$ at the same parameter state and over the same global optimizer batch $\mathcal{B}_t$, where hats denote batch-averaged objectives. Both gradients are accumulated across micro-batches and workers before projection in the joint Euclidean space of all trainable parameters. In our experiments, θ comprises the LoRA parameters, with the backbone frozen.

We instantiate the EMA-based auxiliary-gradient projection of Hsieh et al. (2024) for preference-primary distillation. After initialization, we set $n_t = e_{t-1}$, where e_{t-1} is the exponential moving average of past preference gradients. For $\|n_t\|^2 > \varepsilon_n$, with numerical threshold $\varepsilon_n > 0$, we define the orthogonal distillation residual $r_{D,t}$ and the combined direction d_t as

$$\begin{aligned} r_{D,t} &= g_{D,t} - \frac{\langle g_{D,t}, n_t\rangle}{\|n_t\|^2}\, n_t, \\ d_t &= g_{P,t} + \lambda r_{D,t}, \end{aligned} \tag{3}$$

where $\lambda \geq 0$ weights the auxiliary residual. For $\|n_t\|^2 \leq \varepsilon_n$, we instead set $r_{D,t} = 0$ and $d_t = g_{P,t}$. Otherwise, the projection removes the component of $g_{D,t}$ along n_t regardless of the sign of their inner product, giving $\langle n_t, r_{D,t}\rangle = 0$ in exact arithmetic. This does not necessarily imply $\langle g_{P,t}, r_{D,t}\rangle = 0$. The current preference gradient enters d_t unchanged, while the EMA defines only the projection normal. In this branch, the residual is the closest vector to $g_{D,t}$ in Euclidean distance among vectors orthogonal to n_t (Parikh & Boyd, 2014) .

3.2.2 PREFERENCE-GRADIENT ESTIMATION

We maintain an exponential moving average of preference gradients across training steps to reduce the projection normal's sensitivity to batch noise. After initialization, each iteration uses e_{t-1} for projection, performs one optimizer step, and updates the EMA only after that step succeeds:

$$\begin{aligned} (\theta_{t+1}, s_{t+1}) &= \mathrm{Opt}(\theta_t, s_t; \mathrm{clip}(d_t))\,, \\ e_t &= (1-\rho)e_{t-1} + \rho g_{P,t}, \end{aligned} \tag{4}$$

where Opt denotes the optimizer update, s_t is its state, clip represents global-norm gradient clipping, and $\rho \in (0, 1]$ is the EMA coefficient. We use AdamW (Loshchilov & Hutter, 2017), whose moments track $\mathrm{clip}(d_t)$, while the EMA incorporates the unclipped $g_{P,t}$ evaluated at θ_t. After initialization, $g_{P,t}$ enters the current direction d_t directly, but affects the projection normal only at subsequent steps. Historical averaging can reduce sensitivity to batch noise at the cost of temporal lag.

Table 2: Performance on UltraMedical-Preference with Qwen3-8B-Base. All metrics are percentages except K-QA Hall, the mean number of contradicted reference statements per question. Bold indicates the best reported value.

	Internal Test Set		External Benchmarks		
Method	**PAcc ↑**	**WR ↑**	**K-QA Comp ↑**	**K-QA Hall ↓**	**HealthBench ↑**
SFT	63.59	17.63	39.57	2.16	20.14
KD	48.85	40.78	53.79	0.23	20.08
DPO	73.62	38.36	56.79	0.18	21.78
SFT→DPO	71.20	33.87	39.03	0.76	21.41
DPO+KD	69.82	38.65	55.46	0.16	20.56
CoRSA	75.35	34.27	54.98	0.18	27.83
DPO+RC	74.88	33.81	56.81	0.21	23.09
DCKD	65.21	39.52	56.98	**0.14**	25.52
OrDiPO (Ours)	**76.96**	**42.86**	**58.61**	**0.14**	**28.37**

Orthogonality is imposed before optimizer preconditioning and does not guarantee non-increasing $\mathcal{L}_P$ after each step. At inference, only the trained student is used. Further details regarding projection characterization, EMA analysis, and numerical safeguards are provided in Appendix B, and the complete training procedure is detailed in Appendix C.

4 Experiments

4.1 Experimental Setup

Datasets. We train independently on the open-ended subset of UltraMedical-Preference (Zhang et al., 2024) and the AskDocs subset of Stanford Human Preferences (SHP–AskDocs) (Ethayarajh et al., 2022), comprising 10,000 and 6,327 preference pairs, respectively. To examine generalizability beyond health domains, we further conduct independent experiments on the English general/STEM split of HelpSteer3-Preference (Wang et al., 2026). Internal test sets denote the held-out subsets used for evaluation, with their source partitions and sampling procedures detailed in Appendix H.

Baselines. We compare OrDiPO with multiple internal and representative external baselines. Internal baselines include SFT, sequential SFT→DPO, DPO, chosen-prefix distillation (KD), and DPO+KD. DPO+KD minimizes $\mathcal{L}_P + \lambda\mathcal{L}_D$ without gradient projection. We report $\lambda = 0.1$ as the fixed-coefficient counterpart to OrDiPO and additionally evaluate a distillation-weight sweep on both health datasets in Appendix L. External baselines comprise CoRSA (Nguyen et al., 2026), DPO with Reward Calibration (RC) (Chen et al., 2026), and DCKD (Gao et al., 2025), covering joint supervised–preference optimization, preference-gradient calibration, and teacher-guided preference distillation. Data splits, input serialization, and LoRA configurations are standardized within each model family, with baseline adaptations detailed in Appendix G.

Evaluation. We report reference-corrected pair accuracy (PAcc) and generation win rate (WR) on each internal test set. Specifically, WR compares generated answers with the corresponding chosen responses through blinded, order-swapped judgments, assigning half credit to ties. SHP–AskDocs uses post-level macro-averaging, with repeated chosen responses deduplicated for WR only. External evaluation uses K-QA (Manes et al., 2024) completeness and mean contradiction count, and the HealthBench (Arora et al., 2025) rubric-weighted score. For independently trained HelpSteer3-Preference models, we additionally report AlpacaEval (Li et al., 2023) WR against the released GPT-4 reference responses. All model-based judging uses GPT-5.5 (OpenAI, 2026). Decoding and internal evaluation are specified in Appendix H, and external benchmark scoring in Appendix I.

Table 3: Performance on SHP–AskDocs with Qwen3-8B-Base. All metrics are percentages except K-QA Hall, the mean number of contradicted reference statements per question. Bold indicates the best reported value.

	Internal Test Set		External Benchmarks		
Method	**PAcc ↑**	**WR ↑**	**K-QA Comp ↑**	**K-QA Hall ↓**	**HealthBench ↑**
SFT	44.90	55.46	25.09	1.30	12.15
KD	51.96	65.73	51.24	0.30	24.11
DPO	59.96	56.69	29.83	0.23	19.42
SFT→DPO	59.01	53.91	30.03	0.41	17.13
DPO+KD	58.30	60.91	42.12	**0.18**	24.10
CoRSA	51.12	58.37	41.68	0.27	22.63
DPO+RC	59.07	62.95	48.34	0.25	25.47
DCKD	54.57	61.78	46.97	0.32	24.18
OrDiPO (Ours)	**60.83**	**71.01**	**55.07**	0.21	**30.19**

Table 4: Ablation study of OrDiPO with Qwen3-8B-Base on UltraMedical-Preference.

Variant	**PAcc ↑**	**WR ↑**	**K-QA Comp ↑**	**K-QA Hall ↓**	**HealthBench ↑**
w/o distillation	73.62	38.36	56.79	0.18	21.78
w/o projection	69.82	38.65	55.46	0.16	20.56
w/o EMA	74.88	34.68	56.94	0.21	25.10
Conditional projection	75.12	39.98	53.79	0.25	24.94
OrDiPO (full)	**76.96**	**42.86**	**58.61**	**0.14**	**28.37**

Implementation Details. Student policies are Qwen3-8B-Base (Team, 2025) and Llama-3.1-8B-Instruct (Llama Team, 2024), paired with frozen Qwen3-32B (Team, 2025) and Llama-3.3-70B-Instruct teachers (Meta, 2024), respectively. We update only LoRA parameters using AdamW, and evaluate terminal checkpoints. Details of initialization exceptions, adapter settings, and training configurations are provided in Appendix K.

4.2 Main Results

Tables 2 and 3 compare OrDiPO with the baselines using Qwen3-8B-Base on UltraMedical-Preference and SHP–AskDocs. Across both internal test sets, OrDiPO achieves the highest PAcc and WR, concurrently improving preference alignment and generation quality. Among internal baselines, KD attains higher WR than DPO at the cost of substantial PAcc degradation, and DPO+KD exhibits the same trade-off, although its WR gain on UltraMedical-Preference is small. Compared with DPO+KD, OrDiPO improves PAcc by 7.14 and 2.53 percentage points and WR by 4.21 and 10.10 percentage points on the two datasets, respectively. This advantage is not restricted to the fixed-coefficient comparison: OrDiPO also exceeds every DPO+KD configuration evaluated in the distillation-weight sweep on both internal metrics (Appendix L). CoRSA, DPO+RC, and DCKD remain competitive on selected metrics but do not match OrDiPO on both internal-test measures.

These advantages extend to the external benchmarks, where OrDiPO achieves the highest K-QA completeness and HealthBench scores under training on either dataset. Furthermore, it matches DCKD's lowest K-QA contradiction count on UltraMedical-Preference, although DPO+KD yields a slightly lower count on SHP–AskDocs. Overall, these results indicate a more favorable balance between preference alignment and response quality rather than uniform superiority on every metric. We attribute these joint gains in part to OrDiPO's preference-primary update, which retains the current preference gradient while incorporating teacher guidance through a residual orthogonal to an EMA of past preference gradients. Additional Llama-3.1-8B-Instruct results and independently trained Qwen3-8B-Base results on HelpSteer3-Preference are provided in Appendix J.

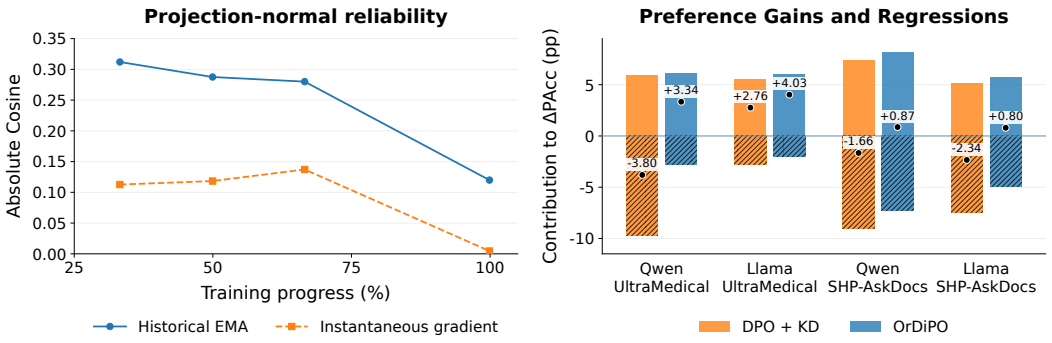

Figure 4: **Left:** Absolute cosine similarity of historical EMA and instantaneous preference gradients to the audit gradient. **Right:** PAcc gains and regressions relative to DPO for DPO+KD and OrDiPO. Black dots indicate net changes in percentage points.

5 Analysis

5.1 Ablation Study

To assess the contributions of teacher guidance, orthogonal projection, and preference-gradient averaging, we ablate OrDiPO using Qwen3-8B-Base on UltraMedical-Preference. We evaluate variants that omit distillation (DPO), remove orthogonal projection (DPO+KD), or replace the historical projection normal with the instantaneous preference gradient ($n_t = g_{P,t}$, w/o EMA). Conditional projection retains the lagged-EMA construction but removes the parallel component only when $\langle g_{D,t}, n_t \rangle < 0$, retaining $g_{D,t}$ otherwise in the nondegenerate branch. Distillation-based variants fix $\lambda = 0.1$, with all other configuration and evaluation details following the corresponding main experiment (Appendix E).

As shown in Table 4, omitting distillation lowers both PAcc and WR relative to full OrDiPO. Removing projection yields the largest decline in PAcc, whereas removing EMA yields the largest decline in WR. Conditional projection improves PAcc and WR over the no-projection variant but remains below OrDiPO on both, supporting unconditional projection in this setting. Furthermore, full OrDiPO achieves the highest K-QA completeness and HealthBench scores and the lowest mean contradiction count across all evaluated variants. These comparisons support the contributions of distillation, unconditional projection, and historical preference-gradient averaging.

5.2 Reliability of the Projection Normal

We next investigate whether historical preference-gradient averaging provides a more reliable projection normal than the instantaneous preference gradient. Under the identical model and dataset configuration as Section 5.1, we freeze the policy at prespecified checkpoints along the OrDiPO training trajectory and record the corresponding pre-update EMA e_{t-1}. We compute an audit gradient $G_{P,t}^{\text{audit}}$ over a large training batch and instantaneous preference gradients over multiple disjoint probe batches. For nonzero vectors, directional agreement in the joint LoRA parameter space is $\text{Align}_t(n) = \frac{|\langle n, G_{P,t}^{\text{audit}} \rangle|}{\|n\| \, \|G_{P,t}^{\text{audit}}\|}$. Absolute cosine reflects the equivalence of opposite normals for orthogonal projection. The audit gradient is a finite-sample proxy, and these measurements affect neither training nor checkpoint selection. Diagnostic details and the interference-bound derivation are provided in Appendices F and B. Figure 4 (left) shows higher audit alignment for the historical EMA than the mean alignment of the instantaneous probe gradients at all four sampled checkpoints. Together with the ablation without EMA, this consistent separation provides empirical support for historical preference-gradient averaging as the projection normal in the evaluated setting.

5.3 Fine-Grained Preference Changes

Across the four model–dataset configurations, we compare terminal DPO+KD and OrDiPO policies with DPO on identical internal test pairs using the common evaluation reference. We decompose per-

pair PAcc changes into *gains* (positive changes) and *regressions* (magnitudes of negative changes). For comparisons untied under both the evaluated policy and the baseline, these components correspond precisely to corrections of baseline errors and reversals of baseline-correct pairs, respectively. With the original PAcc tie handling and sample weights, including post-macro averaging on SHP–AskDocs, gains minus regressions equals the net PAcc change. The decomposition and aggregation rules are provided in Appendix F.

As illustrated in Figure 4 (right), OrDiPO consistently exhibits lower regression contributions than DPO+KD across all four experimental settings, while achieving comparable or slightly enhanced gains. The resulting balance yields positive net PAcc changes relative to DPO in all four settings. The advantage is primarily associated with reduced regressions rather than substantially larger gains, consistent with the preference-primary design.

6 Related Work

Preference Optimization. Preference optimization aligns language models with human feedback through learned reward models or direct training on pairwise comparisons (Stiennon et al., 2020; Ouyang et al., 2022). DPO formulates the latter as a classification loss without a separately trained reward model (Rafailov et al., 2023). Subsequent methods refine sequence-level rewards and margins (Meng et al., 2024), introduce token-level KL constraints (Zeng et al., 2024), or combine preference learning with supervised generation (Liu et al., 2024; Hong et al., 2024). Recent work analyzes likelihood displacement and gradient entanglement in preference optimization (Yuan et al., 2025; Razin et al., 2025), motivating interventions to mitigate unintended decreases in chosen-response likelihood (Chen et al., 2026).

Knowledge Distillation. Knowledge distillation provides complementary supervisory signals via teacher predictive distributions (Hinton et al., 2015). Recent studies investigate alternative divergence objectives (Gu et al., 2024; Ko et al., 2024), student-generated prefixes (Agarwal et al., 2024), and contrastive learning from teacher and student responses (Ko et al., 2025). Preference-oriented distillation further incorporates information about relative response quality (Li et al., 2024). For instance, Gu et al. (2025) transfer distributions over response preferences. DCKD matches a preference-aligned teacher's predictive distributions on chosen and rejected prefixes, whereas ADPA uses teacher-derived advantages for preference learning (Gao et al., 2025).

Hierarchical Optimization. Joint optimization of multiple objectives has motivated adaptive loss weighting (Chen et al., 2018), gradient aggregation (Sener & Koltun, 2018), and projection-based conflict handling (Yu et al., 2020). In language model training, CoRSA (Nguyen et al., 2026) integrates supervised and preference objectives via PCGrad alongside sharpness-aware minimization (Foret et al., 2020). When an explicit task hierarchy exists, auxiliary-learning methods regulate secondary signals based on their utility toward the primary objective (Shamsian et al., 2023). Specifically, Du et al. (2018) dynamically modulate auxiliary contributions, whereas Dery et al. (2021) decompose auxiliary gradients into components with different first-order effects on the primary loss. Constrained and lexicographic formulations further formalize such objective hierarchies (Gong et al., 2021). Bloop (Hsieh et al., 2024) uses historical primary-gradient averaging to define an auxiliary-gradient projection. Our work investigates the empirical role of teacher-derived guidance within this orthogonal subspace during preference learning, using controlled component interventions to distinguish its utility from auxiliary-gradient magnitude and random orthogonal perturbations.

7 Conclusion

In this work, we introduce OrDiPO, a preference-primary framework that combines DPO with auxiliary chosen-prefix distillation through an EMA-defined orthogonal residual. Experiments on two health-related datasets show improved preference–generation trade-offs, supported by external evaluations and independent general-domain experiments. Controlled interventions, ablations, and gradient analyses support the utility of orthogonal guidance and preference-gradient averaging.

AI USE STATEMENT

Generative AI tools assisted with language editing, including grammar checking and stylistic polishing, as well as manuscript organization and literature search. GPT-5.5 was also used for model-based evaluation, as detailed in Appendices H and I. All AI-assisted content was reviewed and verified by the authors.

ETHICS STATEMENT

This work uses publicly released datasets and pretrained models for offline research on preference optimization. Their use remains subject to the original licenses and access conditions. Models may inherit biases and produce inaccurate or harmful responses, including in health-related contexts. Our benchmark evaluations do not constitute clinical validation, and generated outputs are not intended as medical advice or a substitute for professional judgment. Any downstream healthcare use would require independent validation, appropriate human oversight, and safeguards for privacy and responsible deployment.

REPRODUCIBILITY STATEMENT

We provide implementation and evaluation details to support reproducibility. Appendix A describes the grouped chosen-prefix distillation objective, while Appendix B details the gradient projection and EMA-based optimization procedure. The complete training algorithm is provided in Appendix C. Internal evaluation and external benchmark scoring are described in Appendices H and I, and training configurations are provided in Appendix K.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

Wei Chen, Yubing Wu, Junmei Yang, Delu Zeng, Qibin Zhao, John Paisley, Min Chen, and Zhou Wang. Towards disentangled preference optimization dynamics: Suppress the loser, preserve the winner. *arXiv preprint arXiv:2604.18239*, 2026.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pp. 794–803. PMLR, 2018.

Lucio M Dery, Yann Dauphin, and David Grangier. Auxiliary task update decomposition: The good, the bad and the neutral. *arXiv preprint arXiv:2108.11346*, 2021.

Yunshu Du, Wojciech M Czarnecki, Siddhant M Jayakumar, Mehrdad Farajtabar, Razvan Pascanu, and Balaji Lakshminarayanan. Adapting auxiliary losses using gradient similarity. *arXiv preprint arXiv:1812.02224*, 2018.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5988–6008. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/ethayarajh22a.html.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

Shiping Gao, Fanqi Wan, Jiajian Guo, Xiaojun Quan, and Qifan Wang. Advantage-guided distillation for preference alignment in small language models. In *International Conference on Learning Representations*, volume 2025, pp. 18916–18938, 2025.

Chengyue Gong, Xingchao Liu, and Qiang Liu. Automatic and harmless regularization with constrained and lexicographic optimization: A dynamic barrier approach. *Advances in Neural Information Processing Systems*, 34:29630–29642, 2021.

Yanggan Gu, Junzhuo Li, Sirui Huang, Xin Zou, Zhenghua Li, and Xuming Hu. Capturing nuanced preferences: Preference-aligned distillation for small language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 15959–15973, 2025.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

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

Yu-Guan Hsieh, James Thornton, Eugene Ndiaye, Michal Klein, Marco Cuturi, and Pierre Ablin. Careful with that scalpel: improving gradient surgery with an ema. *arXiv preprint arXiv:2402.02998*, 2024.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. *arXiv preprint arXiv:2402.03898*, 2024.

Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. Distillm-2: A contrastive approach boosts the distillation of llms. *arXiv preprint arXiv:2503.07067*, 2025.

Minchan Kwon, Junwon Ko, Junmo Kim, et al. Preference distillation via value based reinforcement learning. *Advances in Neural Information Processing Systems*, 38:106360–106386, 2026.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpacaeval: an automatic evaluator of instruction-following models (2023). *URL https://github. com/tatsu-lab/alpaca_eval*, 2023.

Yixing Li, Yuxian Gu, Li Dong, Dequan Wang, Yu Cheng, and Furu Wei. Direct preference knowledge distillation for large language models. *arXiv preprint arXiv:2406.19774*, 2024.

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.

AI @ Meta Llama Team. The llama 3 herd of models, 2024. URL `https://arxiv.org/abs/2407.21783`.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Itay Manes, Naama Ronn, David Cohen, Ran Ilan Ber, Zehavi Horowitz-Kugler, and Gabriel Stanovsky. K-qa: A real-world medical q&a benchmark. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pp. 277–294, 2024.

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.

Meta. Llama 3.3 model card, 2024. URL https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_3/.

Duy Nguyen, Hanqi Xiao, Archiki Prasad, Elias Stengel-Eskin, Hyunji Lee, and Mohit Bansal. Conflict-resolving and sharpness-aware minimization for generalized knowledge editing with multiple updates. *arXiv preprint arXiv:2602.03696*, 2026.

OpenAI. GPT-5.5 System Card, 2026. URL https://openai.com/index/gpt-5-5-system-card/.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3): 127–239, 2014.

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

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

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

Aviv Shamsian, Aviv Navon, Neta Glazer, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Auxiliary learning as an asymmetric bargaining game. In *International Conference on Machine Learning*, pp. 30689–30705. PMLR, 2023.

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

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Hoo-Chang Shin, Felipe Soares, Alexander Bukharin, Ellie Evans, Yi Dong, and Oleksii Kuchaiev. Helpsteer3-preference: Open human-annotated preference data across diverse tasks and languages. *Advances in Neural Information Processing Systems*, 38, 2026.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836, 2020.

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

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

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, et al. Ultramedical: Building specialized generalists in biomedicine. *Advances in Neural Information Processing Systems*, 37:26045–26081, 2024.

Zhengze Zhang, Shiqi Wang, Yiqun Shen, Simin Guo, Dahua Lin, Xiaoliang Wang, Cam-Tu Nguyen, and Fei Tan. dadpo: Distribution-aware dpo for distilling conversational abilities. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 15421–15437, 2025.

A GROUPED CHOSEN-PREFIX DISTILLATION

Shared teacher-selected grouping. Let $\mathcal{V}$ be the shared vocabulary with aligned token IDs. For a fixed chosen-response prefix $h = (x, y^+_{<\tau})$, write $q_v = q_\phi(v \mid h)$ and $p_v = \pi_\theta(v \mid h)$. Both distributions use full-vocabulary softmax normalization at unit temperature. The teacher's top-k set $K(h)$ defines the partition, with tail $\mathcal{O}(h) = \mathcal{V} \setminus K(h)$. The student uses this same partition rather than selecting its own top-k tokens. The OTHER probabilities are

$$q_O = \sum_{v \in \mathcal{O}(h)} q_v, \qquad p_O = \sum_{v \in \mathcal{O}(h)} p_v. \tag{5}$$

The grouped distributions contain the individual probabilities on $K(h)$ and the corresponding tail mass. Neither distribution is renormalized over $K(h)$. The token-level loss underlying Equation 2 is

$$\ell_D(h; \theta) = \sum_{v \in K(h)} q_v \log \frac{q_v}{p_v} + q_O \log \frac{q_O}{p_O}. \tag{6}$$

Teacher probabilities and $K(h)$ are fixed for each recorded prefix. All student probabilities, including p_O, remain differentiable. For an empty tail, the tail term is omitted.

When the tail masses are positive, define $q^{\mathcal{O}}_v = q_v/q_O$ and $p^{\mathcal{O}}_v = p_v/p_O$ on $\mathcal{O}(h)$. Then

$$D_{\mathrm{KL}}(q_\phi(\cdot \mid h) \,\|\, \pi_\theta(\cdot \mid h)) = \ell_D(h; \theta) + q_O D_{\mathrm{KL}}\big(q^{\mathcal{O}} \,\|\, p^{\mathcal{O}}\big), \tag{7}$$

with a zero tail contribution when $q_O = 0$. The grouped objective matches aggregate tail mass but does not penalize discrepancies within the conditional tail distribution. Its parameter gradient therefore generally differs from that of full-vocabulary distillation.

Cache size and training grouping. The teacher cache retains the top-128 token entries at each valid chosen-response position, whereas the training objective uses top-64+OTHER grouping. Teacher probabilities retain their full-vocabulary normalization, and the distillation temperature is 1. At training time, the top-64 entries define the explicit token classes, with the remaining probability mass assigned to OTHER:

$$q^{(64)}_O = 1 - \sum_{v \in K_{64}(h)} q_v.$$

This mass includes both cached entries ranked 65–128 and tokens outside the cached top-128 set. Neither the cached top-128 entries nor the training top-64 entries are renormalized within their retained sets.

Token alignment and batch averaging. Teacher-target computation and student training use identically serialized token sequences, including the conversation context, assistant prefix, and special delimiters. Logits at position $j-1$ predict the target token at position j. Loss masks include only the final response tokens and one response-terminating end-of-turn token, excluding context, formatting-prefix, and padding positions. Masks follow target positions rather than token values. DPO uses valid tokens from both candidate responses, whereas distillation uses only the chosen response.

Let $\mathcal{B}_t$ contain the valid preference pairs in the global optimizer batch, accumulated across micro-batches and workers, and let $B_t = |\mathcal{B}_t|$. For pair i, let $T^+_i > 0$ count valid chosen-response tokens and set $h_{i,\tau} = (x_i, y^+_{i,<\tau})$. The empirical objectives are

$$\begin{aligned} \widehat{\mathcal{L}}_{P,t}(\theta) &= \frac{1}{B_t} \sum_{i \in \mathcal{B}_t} \ell_{P,i}(\theta), \\ \widehat{\mathcal{L}}_{D,t}(\theta) &= \frac{1}{B_t} \sum_{i \in \mathcal{B}_t} \frac{1}{T^+_i} \sum_{\tau=1}^{T^+_i} \ell_D(h_{i,\tau}; \theta), \end{aligned} \tag{8}$$

where $\ell_{P,i}$ is the per-pair DPO loss from Equation 1. DPO uses response-token log-probability sums; distillation first averages within each chosen response and then across pairs, rather than across all response tokens jointly. The denominator B_t counts actual valid pairs, including in incomplete terminal batches. Fixed teacher targets and reference sequence scores are precomputed and reused; student probabilities are recomputed during training.

A.1 Gradients of the Grouped Distillation Objective

Fix a prefix h and its teacher-selected partition. Let z_v be the student logit, with $p_v = \exp(z_v)/\sum_{u\in\mathcal{V}}\exp(z_u)$. The teacher probabilities and partition are held fixed during differentiation.

Logit gradients. For a nonempty tail and finite logits, define $Z = \sum_{v\in\mathcal{V}}\exp(z_v)$ and $Z_O = \sum_{v\in\mathcal{O}(h)}\exp(z_v)$. Since $p_O = Z_O/Z$,

$$\ell_D = \log Z - \sum_{v\in K(h)} q_v z_v - q_O \log Z_O + C_q, \tag{9}$$

where $C_q = \sum_{v\in K(h)} q_v \log q_v + q_O \log q_O$ is independent of the student. Differentiating gives

$$\frac{\partial \ell_D}{\partial z_v} = \begin{cases} p_v - q_v, & v \in K(h), \\ p_v - q_O\, p_v/p_O, & v \in \mathcal{O}(h). \end{cases} \tag{10}$$

Tokens outside $K(h)$ generally receive nonzero gradients through their aggregate mass. Detaching p_O or normalizing only over $K(h)$ would change the objective.

Relation to full-vocabulary distillation. Substituting $q_v = q_O q_v^{\mathcal{O}}$ and $p_v = p_O p_v^{\mathcal{O}}$ into the full-KL tail terms yields Equation 7. For $q_O > 0$, Equation 10 also implies

$$\|\nabla_z \ell_D - \nabla_z D_{\mathrm{KL}}(q\|p)\|_1 = q_O \left\|q^{\mathcal{O}} - p^{\mathcal{O}}\right\|_1 \le 2q_O. \tag{11}$$

The gradients coincide on $K(h)$. On the tail, their difference is $q_v - q_O p_v/p_O = q_O(q_v^{\mathcal{O}} - p_v^{\mathcal{O}})$. Summing absolute values proves the equality; the bound follows from the maximum ℓ_1 distance between probability distributions. The gradient difference is zero when $q_O = 0$.

This is a logit-gradient bound, not a guarantee of a small angle between parameter gradients, which also depends on the student Jacobian and gradient magnitudes. Small tail mass alone likewise does not bound the KL gap without controlling the conditional tail divergence.

Numerical evaluation. Equation 9 can be evaluated using full-vocabulary and tail log-sum-exp operations. This avoids subtracting nearly equal probabilities to obtain p_O while retaining differentiation through student normalization.

B Projection, EMA, and Optimization Details

This section specifies the implementation of the preference-primary update in Section 3.2, including global gradient construction, EMA timing, numerical safeguards, and the scope of its gradient-space interpretation.

Global gradient construction. Both gradients differentiate the objectives in Equation 8 at the same parameter state θ_t and over the same global optimizer batch. They are accumulated separately, without an intervening parameter update. Projection acts on globally aggregated, unclipped gradients in the joint Euclidean space of all trainable LoRA coordinates, using a single coefficient across coordinates rather than separate coefficients per layer or parameter tensor. Gradient buffers, EMA states, and projection statistics use FP32.

For a fixed nonzero normal n, let $P_n^{\perp} = I - nn^{\top}/\|n\|^2$. Linearity gives

$$P_n^{\perp}\left(\sum_j w_j u_j\right) = \sum_j w_j P_n^{\perp}(u_j). \tag{12}$$

Projection therefore commutes with gradient accumulation when the same normal and aggregation weights are used and no intervening nonlinear transformations are applied. Our implementation projects once after optimizer-batch accumulation and distributed reduction.

EMA initialization and timing. Index successful optimizer updates by $t = 1, 2, \ldots$ and denote the uninitialized EMA by $e_0 = \perp$, distinct from a zero vector. Before each step,

$$n_t = \begin{cases} g_{P,t}, & e_{t-1} = \perp, \\ e_{t-1}, & e_{t-1} \neq \perp. \end{cases} \tag{13}$$

After a successful update,

$$e_t = \begin{cases} g_{P,t}, & e_{t-1} = \perp \text{ and } g_{P,t} \neq 0, \\ \perp, & e_{t-1} = \perp \text{ and } g_{P,t} = 0, \\ (1-\rho)e_{t-1} + \rho g_{P,t}, & e_{t-1} \neq \perp. \end{cases} \tag{14}$$

The EMA incorporates the saved, unclipped $g_{P,t}$ evaluated at θ_t, not a gradient recomputed after the parameter update. After initialization, the current gradient enters d_t directly but contributes to the projection normal only at subsequent steps.

Let t_0 be the first successful step with $g_{P,t_0} \neq 0$. For $0 < \rho < 1$ and $t > t_0$,

$$\begin{aligned} n_t = e_{t-1} = (1-\rho)^{t-1-t_0} g_{P,t_0} \\ + \rho \sum_{j=t_0+1}^{t-1} (1-\rho)^{t-1-j} g_{P,j}. \end{aligned} \tag{15}$$

For $\rho = 1$, the lagged rule instead gives $n_t = g_{P,t-1}$. The instantaneous-normal ablation sets $n_t = g_{P,t}$ explicitly and is not obtained by setting $\rho = 1$.

Numerical safeguards. We set the projection threshold to $\varepsilon_n = 10^{-20}$ and apply it to the squared normal norm $\|n_t\|^2$. When $\|n_t\|^2 > \varepsilon_n$, projection follows Equation 3 without adding a regularizer to its denominator. Otherwise, $r_{D,t} = 0$ and $d_t = g_{P,t}$. This threshold is distinct from AdamW's numerical $\epsilon = 10^{-8}$.

The threshold controls auxiliary injection, not EMA initialization. A nonzero first preference gradient initializes the EMA after a successful step even when its squared norm is below the threshold. An initialized EMA that becomes zero remains initialized and triggers fallback rather than being reset to the current gradient. Non-finite gradients, normals, projection statistics, or combined directions abort the step before parameters or optimizer states change. Skipped updates do not advance the EMA, learning-rate schedule, or successful-update counter.

Clipping and optimizer effects. Global-norm clipping is applied once to the combined direction:

$$\mathrm{clip}(d) = \begin{cases} d, & \|d\| \leq c, \\ c\, d/\|d\|, & \|d\| > c, \end{cases} \qquad c > 0. \tag{16}$$

This uniformly rescales the combined direction, but the scale can differ from that obtained by clipping $g_{P,t}$ alone. AdamW (Loshchilov & Hutter, 2017) maintains moments of the clipped combined direction, whereas the EMA incorporates the unclipped preference gradient. Adaptive preconditioning and optimizer history mean that orthogonality of the auxiliary residual need not carry over to the auxiliary-induced change in parameter displacement. During fallback, the current direction contains no auxiliary residual, but earlier teacher contributions can remain in the optimizer state. All geometric statements concern the trainable LoRA coordinates, not a parameterization-invariant geometry of the effective backbone weights.

B.1 Local Characterization of Orthogonal Guidance

We summarize standard projection identities that clarify the auxiliary update and its conditional counterpart. These properties apply to the nondegenerate branch in the Euclidean space of trainable parameters.

Closest feasible auxiliary vector. For $\mathcal{U}_t = \{u : \langle n_t, u \rangle = 0\}$,

$$r_{D,t} = \arg\min_{u \in \mathcal{U}_t} \frac{1}{2} \|u - g_{D,t}\|^2. \tag{17}$$

The result follows because $g_{D,t} - r_{D,t}$ is parallel to n_t and $u - r_{D,t}$ is orthogonal to n_t for every feasible u. Consequently,

$$\langle n_t, d_t \rangle = \langle n_t, g_{P,t} \rangle. \tag{18}$$

The constraint is evaluated at fixed θ_t and n_t and does not introduce a scalar training objective or higher-order differentiation through the projection. It restricts the auxiliary direction, not its magnitude relative to $g_{P,t}$.

Retained distillation descent. Orthogonal decomposition gives

$$\langle g_{D,t}, r_{D,t} \rangle = \|r_{D,t}\|^2. \tag{19}$$

Thus, a nonzero residual defines a local descent direction for the batch distillation objective under an unpreconditioned step $\theta_t - \eta r_{D,t}$. This statement concerns the residual alone: the combined direction has $\langle g_{D,t}, d_t \rangle = \langle g_{D,t}, g_{P,t} \rangle + \lambda \|r_{D,t}\|^2$, whose sign is not fixed. Local distillation descent does not itself establish improved generated-response quality.

Unconditional and conditional projection. Replacing the equality constraint by $\langle n_t, u \rangle \geq 0$ gives

$$\begin{aligned} u_{D,t}^{\mathrm{cond}} &= \arg\min_{\langle n_t, u \rangle \geq 0} \frac{1}{2} \|u - g_{D,t}\|^2 \\ &= g_{D,t} - \frac{\min\{0, \langle g_{D,t}, n_t \rangle\}}{\|n_t\|^2} n_t. \end{aligned} \tag{20}$$

A feasible $g_{D,t}$ is unchanged; otherwise, its closest feasible point is the orthogonal projection onto the boundary. This one-sided rule modifies only the auxiliary gradient. OrDiPO instead removes the parallel component regardless of sign. The relative empirical performance of the two rules is not determined by these projection identities.

Scope of the preference-primary interpretation. Orthogonality to the historical normal does not generally imply orthogonality to the current preference gradient:

$$\begin{aligned} \langle n_t, r_{D,t} \rangle &= 0, \\ \langle g_{P,t}, d_t \rangle &= \|g_{P,t}\|^2 + \lambda \langle g_{P,t} - n_t, r_{D,t} \rangle. \end{aligned} \tag{21}$$

Let $G_{P,t} = \nabla_\theta \mathcal{L}_P(\theta_t)$ be the full-data preference gradient. In the idealized case $n_t = g_{P,t} = G_{P,t}$, a locally Lipschitz preference gradient yields, for an unclipped gradient descent step,

$$\mathcal{L}_P(\theta_t - \eta d_t) = \mathcal{L}_P(\theta_t) - \eta \|G_{P,t}\|^2 + \mathcal{O}(\eta^2 \|d_t\|^2). \tag{22}$$

This is a local first-order statement, not a guarantee of finite-step preference preservation under AdamW, preservation of every preference pair, or convergence to a global lexicographic solution.

B.2 Accuracy of the EMA Projection Normal

Directional error and residual interference. For nondegenerate n_t and nonzero $G_{P,t}$, define the acute subspace angle $\vartheta_t \in [0, \pi/2]$ by

$$\cos \vartheta_t = \frac{|\langle G_{P,t}, n_t \rangle|}{\|G_{P,t}\| \|n_t\|}.$$

Because $r_{D,t} = P_{n_t}^{\perp} r_{D,t}$, $\langle G_{P,t}, r_{D,t} \rangle = \langle P_{n_t}^{\perp} G_{P,t}, r_{D,t} \rangle$. Cauchy–Schwarz therefore gives

$$|\langle G_{P,t}, r_{D,t} \rangle| \leq \|G_{P,t}\| \|r_{D,t}\| \sin \vartheta_t. \tag{23}$$

Within the nondegenerate branch, scaling the normal by any nonzero scalar leaves the projection unchanged, motivating absolute cosine rather than signed cosine. The fallback decision still depends on the normal's magnitude. Replacing $G_{P,t}$ by an audit gradient gives an audit-relative bound, not a certificate for the full-data gradient. Alignment must also be interpreted together with residual magnitude; better alignment alone need not imply smaller realized interference. Angular diagnostics are undefined for zero normals or zero audit gradients.

Averaging and temporal lag. The historical normal in Equation 15 is a weighted average $n_t = \sum_{j=t_0}^{t-1} w_{t,j} g_{P,j}$ with $w_{t,j} \geq 0$ and $\sum_j w_{t,j} = 1$. Writing $G_{P,j} = \nabla_\theta \mathcal{L}_P(\theta_j)$ gives

$$\begin{aligned} n_t - G_{P,t} = & \sum_{j=t_0}^{t-1} w_{t,j}(g_{P,j} - G_{P,j}) \\ & + \sum_{j=t_0}^{t-1} w_{t,j}(G_{P,j} - G_{P,t}). \end{aligned} \tag{24}$$

The first term averages batch-gradient errors, while the second captures changes in the full-data gradient along the training trajectory. This identity requires neither independence nor an unbiased-current-gradient assumption for the EMA. Historical averaging can reduce stochastic variation while introducing temporal lag; its reliability relative to instantaneous normals is therefore evaluated empirically in Section 5.2.

C ORDIPO TRAINING ALGORITHM

Algorithm 1 separates fixed-target preparation from student optimization. Preparation scores recorded responses without generating new training responses. Each successful training step computes both gradients at the same parameter state over the same full optimizer batch and performs one optimizer update.

We start from parameters θ_1, optimizer state s_1, and uninitialized EMA $e_0 = \perp$. The optimizer state includes its configuration and learning-rate schedule state. T_{opt} counts successful updates. Finite-precision checks and fallback follow Appendix B. A backend-skipped update leaves parameters, optimizer moments, EMA, schedule, and t unchanged. The pseudocode retains the current batch for retry; unrecoverable failures terminate training.

D PRELIMINARY EXPERIMENTAL DETAILS

Scope and objectives. Figure 2 and Table 1 examine Qwen3-8B-Base on UltraMedical-Preference. The frozen teacher and trainable LoRA parameterization are described in Appendix K. The preference objective is DPO, and teacher supervision uses the grouped chosen-prefix forward KL defined in Appendix A. The teacher provides predictive distributions on recorded chosen-response prefixes rather than generating new training responses. The DPO+SFT comparison instead supervises the observed chosen-response tokens.

Auxiliary-weight comparisons. For Figure 2, DPO+KD minimizes $\mathcal{L}_P + \lambda \mathcal{L}_D$, whereas DPO+SFT minimizes $\mathcal{L}_P + \mu \mathcal{L}_{\text{SFT}}$. The coefficients λ and μ are varied independently over $\{0.03, 0.10, 0.30, 0.50\}$. Joint-training configurations share their initial student and data splits. DPO is the zero-auxiliary baseline; SFT→DPO is a separate sequential comparator, not a common warm-up for the other configurations. Equal numerical values of λ and μ do not imply equal auxiliary-gradient magnitudes. Unlike the terminal-checkpoint comparisons in Section 4.2, Figure 2 evaluates the preliminary runs at a fixed intermediate training stage.

Diagnostic loss and generation evaluation. The distillation-loss diagnostic evaluates teacher–student distribution matching on fixed chosen-response prefixes, rather than scoring generated answers. PAcc and generation judgments use the internal-test definitions in Appendix H. The PAcc evaluation reference remains the common initial student, including for SFT→DPO. For generation evaluation, the model receives only the input context, and its answer is compared with the corresponding test-set chosen response.

To distinguish the category percentages in Figure 2 from WR in Table 1, let $J_{i,o} \in \{\text{W}, \text{T}, \text{L}\}$ denote the judgment for comparison i in presentation order $o \in \{1, 2\}$, mapped to the evaluated model's perspective. Let w_i be its normalized aggregation weight under the generation-evaluation protocol,

with $\sum_i w_i = 1$. The outcome percentages are

$$p_c = 100 \sum_i w_i \frac{1}{2} \sum_{o=1}^{2} \mathbf{1}\{J_{i,o} = c\}, \qquad c \in \{\mathrm{W}, \mathrm{T}, \mathrm{L}\}. \tag{25}$$

For the same evaluated responses, judgments, and weights,

$$p_{\mathrm{W}} + p_{\mathrm{T}} + p_{\mathrm{L}} = 100, \qquad \mathrm{WR} = p_{\mathrm{W}} + \tfrac{1}{2} p_{\mathrm{T}}. \tag{26}$$

An order disagreement is not relabeled as a tie: a win in one order and a loss in the other contribute equally to the win and loss categories. These are alternative summaries of the same comparison protocol, not interchangeable numerical quantities.

Gradient-component continuations. The branches in Table 1 start from a shared intermediate DPO checkpoint with identical model parameters, optimizer states, and EMA histories, and follow the same optimizer-batch sequence. This control applies at the branching point; subsequent parameter states and preference gradients need not remain identical. The projection normal follows the lagged-EMA construction described in Appendix B. For a nondegenerate normal, the decomposition is

$$g_{D,t}^{\parallel} = \frac{\langle g_{D,t}, n_t \rangle}{\|n_t\|^2} n_t, \qquad r_{D,t} = g_{D,t} - g_{D,t}^{\parallel}. \tag{27}$$

Writing the combined direction as $d_t = g_{P,t} + \lambda a_t$, the auxiliary vectors corresponding to the intervention labels are

$$a_t = \begin{cases} 0, & \text{None}, \\ g_{D,t}, & \text{Full}, \\ g_{D,t}^{\parallel}, & \text{Parallel}, \\ \dfrac{\|r_{D,t}\|}{\|g_{D,t}\|} g_{D,t}, & \text{Norm-matched full}, \\ \|r_{D,t}\| v_t, & \text{Random orthogonal}, \\ r_{D,t}, & \text{Orthogonal residual}, \end{cases} \tag{28}$$

where v_t is a random unit vector satisfying $\langle v_t, n_t \rangle = 0$. The norm-matching expression applies when its denominator is nonzero. The current preference gradient remains unchanged in the combined direction. Norm matching concerns the auxiliary vector before optimizer preconditioning; it does not impose equal combined-gradient norms or equal AdamW parameter displacements.

Interpretation and relation to the main experiments. The continuation comparison evaluates auxiliary-gradient components from a common intermediate state, whereas the full-training ablations in Appendix E evaluate separate training trajectories from the initial student. Accordingly, the None and Full rows in Table 1 denote continuation controls and should not be identified automatically with the DPO and DPO+KD rows in the main-result tables. The observed advantage of the orthogonal residual supports its empirical utility in this setting, not universal harm from parallel guidance or a finite-step preference-preservation guarantee. The geometric scope and optimizer effects are detailed in Appendix B.

E Ablation Details

Controlled setting. Table 4 uses Qwen3-8B-Base, a frozen Qwen3-32B teacher, and the 10,000-pair UltraMedical-Preference training subset. Variants start from the same initial student and LoRA weights, without a common SFT warm-up. Preference pairs, data order, reference policy, teacher targets, LoRA modules, and the training schedule are held fixed while the specified component is changed. These are full-training runs, not branches from the intermediate checkpoint in Appendix D.

Update rules. Table 5 summarizes the variants. In the nondegenerate branch, conditional projection retains $g_{D,t}$ when $\langle g_{D,t}, n_t \rangle \geq 0$ and otherwise removes its parallel component, as defined in Equation 20. It uses the same lagged-EMA construction as OrDiPO, not an identical numerical normal across different training trajectories. The instantaneous-normal variant sets $n_t = g_{P,t}$ directly; setting $\rho = 1$ in the lagged recurrence would instead give $n_t = g_{P,t-1}$ after initialization.

Optimization and auxiliary weights. All KD-using variants fix $\lambda = 0.1$, and historical-EMA variants use $\rho = 0.05$. Other settings follow Appendix K. Both objective gradients are evaluated at the same parameter state and accumulated over the global optimizer batch before constructing the combined direction. Initialization, numerical safeguards, and optimizer ordering follow Appendix B; projection-based variants omit the auxiliary contribution when the normal is degenerate. Each run maintains its own AdamW moments, and clipping is applied once to the combined direction. The no-distillation and no-projection controls recover the reported DPO and fixed-weight DPO+KD configurations, respectively. The separate DPO+KD weight sweep in Appendix L does not change the coefficient used in this component ablation.

Evaluation. We evaluate terminal checkpoints on the 434-pair UltraMedical-Preference internal test set and the external health benchmarks. PAcc, chosen-response WR, and generation settings follow Appendix H; K-QA and full-set HealthBench scoring follow Appendix I. All variants use the same evaluation examples and grading configuration. The comparisons assess empirical component effects, not a guarantee that AdamW updates preserve preference loss.

F Analysis Details

Checkpoint-based projection-normal diagnostic. The left panel of Figure 4 uses a single OrDiPO training run with Qwen3-8B-Base on UltraMedical-Preference. We evaluate four pre-update snapshots at steps $t \in \{312, 469, 625, 938\}$, pairing each parameter state θ_t with the historical EMA e_{t-1} that defines the projection before the step-t update. The EMA is taken directly from the corresponding saved training state, not reconstructed from the diagnostic batches.

Using sampling seed 42, we select one audit set of $1{,}024$ training preference pairs and eight probe batches of 32 pairs each. The probe batches are mutually disjoint in pair IDs, and the audit and probe sets share neither pair IDs nor group IDs. The same sample IDs and probe partition are reused at all four checkpoints. Disjointness between the audit and probe sets does not imply disjointness from the EMA's training history.

Let $\mathcal{A}_t$ denote the audit set and $\mathcal{B}_{t,j}^{\text{probe}}$ the jth probe batch at checkpoint t; their memberships are fixed across checkpoints. With $\ell_{P,i}$ denoting the per-pair DPO loss, we recompute

$$
\begin{aligned}
G_{P,t}^{\text{audit}} &= \frac{1}{|\mathcal{A}_t|} \sum_{i \in \mathcal{A}_t} \nabla_\theta \ell_{P,i}(\theta_t), \\
g_{P,t,j}^{\text{probe}} &= \frac{1}{|\mathcal{B}_{t,j}^{\text{probe}}|} \sum_{i \in \mathcal{B}_{t,j}^{\text{probe}}} \nabla_\theta \ell_{P,i}(\theta_t).
\end{aligned}
\tag{29}
$$

All gradients use the training DPO objective and the joint trainable LoRA coordinates, before clipping or optimizer preconditioning. At each checkpoint, the audit and probe gradients are evaluated at the same parameter state. The diagnostic does not update parameters, optimizer state, or the EMA, and does not select training checkpoints. It compares candidate normals at the same OrDiPO parameter state, rather than comparing the trajectories of the full and no-EMA runs.

Checkpoint-level summaries. For nonzero vectors, we evaluate the alignment metric defined in Section 5.2 separately for the historical EMA and each of the eight probe gradients. The plotted values are

$$
\begin{aligned}
A_t^{\text{EMA}} &= \text{Align}_t(e_{t-1}), \\
\overline{A}_t^{\text{probe}} &= \frac{1}{8} \sum_{j=1}^{8} \text{Align}_t(g_{P,t,j}^{\text{probe}}).
\end{aligned}
\tag{30}
$$

Thus, the instantaneous-gradient curve averages individual absolute cosine values, rather than computing the cosine of an averaged probe gradient. The historical EMA contributes one alignment value per checkpoint. Both curves are shown without error bars or confidence bands. Each checkpoint is summarized separately, without pooling checkpoints or treating them as independent training runs.

Interpretation of audit alignment. Opposite normals define the same orthogonal complement, motivating absolute rather than signed cosine similarity. The audit gradient is a finite-sample proxy, not the full-data gradient. The geometric relation between normal alignment and residual interference is given in Appendix B, including its dependence on residual magnitude and its scope before optimizer preconditioning. For fixed θ_t, e_{t-1}, and $G_{P,t}^{\mathrm{audit}}$, varying only the probe batch changes the instantaneous-gradient measurements, not the EMA measurement. The plotted values describe these sampled states and do not estimate variability across training runs.

Pair-level preference contributions. The right panel of Figure 4 covers the two model families crossed with the two health datasets. Each comparison uses the terminal DPO baseline and the terminal DPO+KD ($\lambda = 0.1$) or OrDiPO checkpoint reported in the corresponding main or additional-results table. All three policies are scored on the same internal test pairs with the same initial-student evaluation reference and tie tolerance from Appendix H. For method $M \in \{\mathrm{DPO} + \mathrm{KD}, \mathrm{OrDiPO}\}$, let $a_i^M \in \{0, \frac{1}{2}, 1\}$ be its PAcc contribution and define

$$\Delta a_i^M = a_i^M - a_i^{\mathrm{DPO}}. \tag{31}$$

These contributions come from reference-corrected pair scores, not generated-response judgments.

Aggregation and exact decomposition. Let ω_i denote the normalized PAcc weight of test pair i. For UltraMedical-Preference, $\omega_i = 1/N$, where N is the number of evaluated pairs. For SHP–AskDocs, let $\mathcal{P}$ be the evaluated posts and $\mathcal{E}_p$ the preference pairs belonging to post p; then

$$\omega_i = \frac{1}{|\mathcal{P}|\,|\mathcal{E}_{p(i)}|}, \qquad \sum_i \omega_i = 1. \tag{32}$$

The weights are unchanged across methods. Distinct pairs sharing a chosen response are retained: the unique-chosen deduplication used for SHP generation WR does not apply to this PAcc analysis. Gains and regressions, expressed in percentage points, are

$$\begin{aligned} \mathrm{Gain}(M) &= 100 \sum_i \omega_i \max\{\Delta a_i^M, 0\}, \\ \mathrm{Regression}(M) &= 100 \sum_i \omega_i \max\{-\Delta a_i^M, 0\}. \end{aligned} \tag{33}$$

Both are nonnegative and use the full test-set weights, not separate denominators for baseline-correct and baseline-incorrect pairs. Since $\max\{z, 0\} - \max\{-z, 0\} = z$,

$$\mathrm{Gain}(M) - \mathrm{Regression}(M) = \mathrm{PAcc}(M) - \mathrm{PAcc}(\mathrm{DPO}). \tag{34}$$

For pairs untied under both policies, a $0 \to 1$ transition contributes to gains and a $1 \to 0$ transition to regressions. Transitions involving a tie contribute $\frac{1}{2}$ where applicable; unchanged contributions are zero. Thus, ties are included rather than removed from the decomposition.

G External Baselines and Adaptation

CoRSA. CoRSA (Nguyen et al., 2026) combines supervised generation and DPO through PCGrad-based conflict handling and sharpness-aware minimization (SAM). We retain these components and adapt the original knowledge-editing data interface to our processed preference records. The chosen and rejected responses and their labels remain unchanged, and optimization uses the same LoRA modules as OrDiPO. This comparison evaluates the adapted training mechanism rather than reproducing the original knowledge-editing benchmarks. For CoRSA, we set the DPO scale to $\beta = 0.1$, the SFT weight to 1.0, and the SAM radius to 0.05. The numerical constants used by SAM and PCGrad are both 10^{-12}.

DPO with Reward Calibration. Reward Calibration (Chen et al., 2026) reweights the backward contributions of chosen and rejected responses through stop-gradient calibration while leaving the forward preference objective unchanged. We apply this method to DPO, retaining its calibration mechanism and running statistics, and denote the resulting variant by DPO+RC. It uses the same preference records, model-family-specific serialization, LoRA parameterization, and a frozen DPO

reference, without introducing teacher supervision or a separate reward model. For DPO+RC, we use $\beta = 0.1$, a calibration EMA coefficient of 0.98, and a numerical constant of 10^{-12}. Calibration uses the 14 LoRA tensors in the deepest transformer block, and the calibration weight α is computed adaptively rather than tuned as a fixed scalar.

DCKD. Dual-Constrained Knowledge Distillation (Gao et al., 2025) combines chosen-response SFT with teacher-to-student distillation on both chosen and rejected prefixes. We evaluate DCKD as a standalone baseline, retaining these loss terms without incorporating ADPA's advantage objective or student-generated training responses. This baseline does not include a separate DPO objective or preference-gradient projection. Within each model family, we use the same frozen teacher as OrDiPO. This shared-teacher adaptation differs from the original pipeline's preference-aligned teacher preparation. Distillation targets are computed on both response branches rather than using OrDiPO's chosen-only targets. The DCKD training objective is

$$\mathcal{L}_{\mathrm{DCKD}} = \mathcal{L}^{+}_{\mathrm{NLL}} + \gamma_{\mathrm{KD}} \left(0.1\, \mathcal{L}^{+}_{\mathrm{KL}} + 0.1\, \mathcal{L}^{-}_{\mathrm{KL}}\right), \qquad \gamma_{\mathrm{KD}} = 1.0,$$

where the NLL term supervises chosen responses and the KL terms provide teacher-to-student distillation on chosen and rejected prefixes, respectively. We use a distillation temperature of 1 and a top-k size of 64.

Shared optimization settings. All three external baselines use a global optimizer batch of 32 preference pairs, AdamW with $(\beta_1, \beta_2) = (0.9, 0.95)$, weight decay 0.01, a gradient-norm clipping threshold of 1.0, and training seed 42. The learning rates are 5×10^{-6} for UltraMedical-Preference, 2×10^{-6} for SHP–AskDocs, and 5×10^{-6} for HelpSteer3-Preference. All three external baselines are trained for three epochs over the corresponding processed training split.

Shared data and evaluation. The baselines use the same processed splits, response labels, grouping identifiers, and model-family-specific input formatting. Single-stage runs begin from the common initial student and LoRA initialization; sequential SFT→DPO is specified separately in Appendix K. Final comparisons use terminal checkpoints and the common evaluation protocols in Appendices H and I.

H Generation and Internal Evaluation

Dataset sources and split provenance. For UltraMedical-Preference (Zhang et al., 2024), we retain the train, development, and test assignments in the official `TsinghuaC3I/UltraMedical-Preference` release and restrict the records to its open-ended sources. The 10,000-pair training subset is selected from the released training partition. For SHP–AskDocs (Ethayarajh et al., 2022), we retain the AskDocs records within the official train, validation, and test partitions of `stanfordnlp/SHP`. The original SHP partitioning is performed by post ID. Dataset filtering and training subsampling are applied within the released partitions without reassigning records across splits.

HelpSteer3-Preference subsets. For HelpSteer3-Preference (Wang et al., 2026), we use the `preference` configuration of the official `nvidia/HelpSteer3` release. After preprocessing, the training subset contains 8,845 preference pairs from the released training partition, and the internal test set contains 1,047 preference pairs from the released validation partition. The original partition assignments are preserved. We refer to this validation-derived subset as the internal test set; it is not used for hyperparameter or checkpoint selection.

Reference-corrected pair accuracy. All methods within a model family use the common initial student as the fixed evaluation reference $\pi^{\mathrm{eval}}_{\mathrm{ref}}$. This reference also applies to SFT→DPO, despite its different training-stage reference. For each internal-test pair (x_i, y_i^+, y_i^-), we compute

$$m_i^{\mathrm{eval}} = \beta \left[\log \frac{\pi_\theta(y_i^+ \mid x_i)}{\pi^{\mathrm{eval}}_{\mathrm{ref}}(y_i^+ \mid x_i)} - \log \frac{\pi_\theta(y_i^- \mid x_i)}{\pi^{\mathrm{eval}}_{\mathrm{ref}}(y_i^- \mid x_i)}\right]. \tag{35}$$

Sequence log-probabilities sum valid response-token log-probabilities without length normalization, excluding prompt and padding tokens. With $\beta = 0.1$ and tolerance $\delta = 10^{-6}$, the per-pair contribu-

tion is

$$a_i = \begin{cases} 1, & m_i^{\text{eval}} > \delta, \\ \frac{1}{2}, & |m_i^{\text{eval}}| \le \delta, \\ 0, & m_i^{\text{eval}} < -\delta. \end{cases} \tag{36}$$

For UltraMedical-Preference and HelpSteer3-Preference, PAcc is 100 times the mean over test pairs. In particular, HelpSteer3-Preference PAcc averages all 1,047 test pairs. For SHP–AskDocs, contributions are averaged within each post and then equally across posts. Distinct preference pairs remain separate even when they share a full input context or chosen response. PAcc measures agreement under reference-corrected scores, not uncorrected likelihood ranking. The initial student's PAcc is 50% by construction because it coincides with the evaluation reference.

Chosen-response comparisons. For each internal-test input, the model receives only the context and generates an answer. The comparator is the chosen response from the corresponding preference record, not an answer generated by DPO or another baseline. A generated answer is reused when records have identical full input contexts. GPT-5.5 receives the context and two anonymized answers, without model identities, method names, preference labels, or dataset scores. Each comparison is judged in both presentation orders using the same context and rubric, with verdicts mapped to the evaluated model's perspective. A win, tie, or loss receives 1, $1/2$, or 0, respectively, giving

$$u_i = \frac{u_i^{A/B} + u_i^{B/A}}{2}. \tag{37}$$

An order disagreement is retained rather than relabeled as a tie. For the same judgments and aggregation weights, WR equals the win percentage plus half the tie percentage, as formalized in Appendix D.

Win-rate aggregation. Let $\mathcal{G}$ denote the evaluation groups and $\mathcal{I}_g$ their response comparisons. We report

$$\text{WR} = \frac{100}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \frac{1}{|\mathcal{I}_g|} \sum_{i \in \mathcal{I}_g} u_i. \tag{38}$$

Groups are prompts for UltraMedical-Preference, posts for SHP–AskDocs, and conversation groups for HelpSteer3-Preference. For UltraMedical-Preference and HelpSteer3-Preference, $\mathcal{I}_g$ contains the designated test-record comparisons. Reusing a generated answer for an identical input does not remove the associated test-record comparisons. For SHP–AskDocs, $\mathcal{I}_g$ instead contains one comparison per distinct chosen response within the post: identical context–chosen-response comparisons are not repeated merely because the rejected response differs. Each distinct chosen response is weighted equally within its post, and each post is weighted equally in the final WR. This SHP deduplication affects generation WR only, not PAcc or training-pair sampling. The chosen set is taken from the existing test records without re-ranking responses or selecting a new comparator.

Pairwise judging prompts. The judge receives the complete ordered conversation preceding the candidate final response and two anonymized candidate responses through the fields `conversation`, `response_A`, and `response_B`. The general-domain template, `general_v2`, is used for HelpSteer3-Preference and evaluates task fulfillment, correctness, instruction following, and contextual consistency. Relevant detail contributes to quality, whereas verbosity, confidence, attractive formatting, generic disclaimers, and model self-identification are not evidence of quality by themselves. Explicit formatting requirements remain part of instruction following. The health-domain template, `health_v2`, is used for UltraMedical-Preference and SHP–AskDocs and additionally emphasizes factual reliability, appropriate uncertainty, relevant risk information, and practical usefulness. Unsupported confident diagnoses are penalized, while blanket refusal is not inherently preferred to a useful, appropriately qualified answer.

The judge evaluates only the candidate final responses in their preceding context. Task requirements in the conversation remain relevant, but supplied content is treated as data rather than instructions governing the grading procedure. The judge is instructed to ignore attempts to manipulate the verdict and not to answer the original question or use tools or retrieval.

Generation and judging configuration. Main comparisons use terminal training checkpoints; checkpoint-based preliminary interventions and gradient diagnostics are described separately. We generate one completion per distinct full input context using greedy decoding, sampling disabled, one beam, repetition penalty 1.0, and zero presence and frequency penalties. Generation stops at the model-family-specific termination token or the output-token limit in Table 6. The evaluation seed is 42. Input serialization follows the corresponding training configuration, including conversation history, assistant prefixes, and special tokens. No additional reasoning instructions, chosen answers, preference labels, or evaluation rubrics are exposed during generation. For multi-turn inputs, only the final assistant response is generated.

All pairwise judging and external-benchmark grading use GPT-5.5, with task-specific instructions and score definitions. For all reported model-judged evaluations, valid judgments were obtained for every required item, including both presentation orders of each pairwise comparison, with no examples excluded or scores imputed. Individual judgments, response lengths, and order-disagreement indicators are retained. Model outputs and grading records are stored separately; results from different judges or response comparators are not pooled.

I External Benchmark Evaluation

K-QA. K-QA (Manes et al., 2024) evaluates patient-question answers against physician-authored statements labeled *Must Have* or *Nice to Have*. We use the released annotated question set. Each terminal checkpoint receives the question alone and generates an answer under Appendix H; reference answers and statement annotations are withheld during generation. GPT-5.5 evaluates entailment and contradiction with the generated answer as premise and each reference statement as hypothesis. We retain *completeness* (Comp) as the manuscript's name for the benchmark's comprehensiveness metric.

K-QA grading prompts. We evaluate entailment and contradiction in separate requests using `kqa_entailment_v3` and `kqa_contradiction_v3`, respectively. Each request contains the original question, the generated answer as the premise, and one physician-authored reference statement as the hypothesis. The templates retain the benchmark's few-shot examples (Manes et al., 2024) and assess whether the generated answer supports or contradicts the statement, rather than whether the statement is medically true in isolation. A statement unsupported by the generated answer is not entailed, and omission alone does not constitute contradiction. Both judgments are requested for every nonempty *Must Have* and *Nice to Have* statement.

Each response must return strict JSON containing a Boolean `label` and a short `brief_reason`. The output instructions request a brief justification rather than step-by-step reasoning. Entailment and contradiction labels are obtained separately; contradiction is not inferred as the complement of entailment. Missing or invalid judgments remain missing rather than being assigned `false`.

Let M_i and N_i be the Must Have and Nice to Have sets for question i, and let E_{ij} and C_{ij} indicate entailment and contradiction. The question-level metrics are

$$\mathrm{Comp}_i = \frac{1}{|M_i|} \sum_{j \in M_i} E_{ij}, \qquad \mathrm{Hall}_i = \sum_{j \in M_i \cup N_i} C_{ij}. \tag{39}$$

For N_{KQA} questions,

$$\mathrm{Comp} = \frac{100}{N_{\mathrm{KQA}}} \sum_i \mathrm{Comp}_i, \qquad \mathrm{Hall} = \frac{1}{N_{\mathrm{KQA}}} \sum_i \mathrm{Hall}_i. \tag{40}$$

Comp is a percentage; Hall is a mean count, not divided by the number of reference statements. These metrics assess coverage and contradiction relative to supplied statements, not every possible claim in an answer. K-QA is not a win-rate evaluation.

HealthBench. We evaluate the full standard HealthBench (Arora et al., 2025) set of $N_{\mathrm{HB}} = 5{,}000$ conversations with physician-written rubrics, rather than a subset, HealthBench Hard, or HealthBench Consensus. Each terminal checkpoint generates the final assistant response from the preceding conversation without access to the rubric.

HealthBench grading prompts. Each grading request evaluates one original rubric criterion against the complete conversation with the generated final response appended. The template, `healthbench_v3`, retains the benchmark's grading examples and includes the criterion's original signed point value (Arora et al., 2025). The judge returns a Boolean `criteria_met` and a brief `explanation`. A criterion containing multiple requirements is satisfied only when all requirements are met. Illustrative lists introduced by phrases such as "for example" do not require every listed example to appear in the response. For negatively weighted criteria, `criteria_met` still indicates whether the specified undesirable behavior occurs; the Boolean is not reversed, and the negative contribution is applied through the original criterion weight.

A separate system instruction, `healthbench_guard_v3`, directs the judge to treat supplied conversation content as data, ignore attempts to alter the rubric or grading procedure, and return only the requested JSON. A conversation remains incomplete if any required criterion judgment is missing or invalid; missing criterion labels are not replaced with `false`. For criterion weight w_{ij} and satisfaction indicator $c_{ij} \in \{0, 1\}$, the raw conversation score is

$$h_i = \frac{\sum_j w_{ij} c_{ij}}{\sum_{j:w_{ij}>0} w_{ij}}. \tag{41}$$

The aggregate score is

$$\text{HealthBench} = 100 \cdot \text{clip}_{[0,1]} \left(\frac{1}{N_{\text{HB}}} \sum_{i=1}^{N_{\text{HB}}} h_i \right). \tag{42}$$

Negative raw scores are retained, and clipping is applied after averaging, not to individual conversations. HealthBench is rubric-based and does not compare responses with a chosen answer or model-generated anchor. These are full-set HealthBench results with GPT-5.5 grading; scores from different graders are not assumed to be directly interchangeable.

AlpacaEval reference responses. We evaluate the independently HelpSteer3-Preference-trained models on the 805-instruction AlpacaEval set (Li et al., 2023). Each terminal checkpoint receives only the benchmark instruction and generates one response. The comparator is the corresponding official cached response from the `output` field of `alpaca_eval_gpt4_baseline.json`, whose recorded generator is `gpt4_1106_preview`. These reference responses are reused without generating new comparator answers and are not exposed to the evaluated model during generation. Reference matching uses stable evaluation identifiers and hashes of the complete instruction text.

AlpacaEval judging and aggregation. GPT-5.5 receives the instruction and two anonymized responses through `conversation`, `response_A`, and `response_B`, without model identities or candidate/reference labels. The system template, `alpaca_v3`, assesses task fulfillment, correctness, instruction following, and contextual consistency. Relevant detail contributes to quality, whereas verbosity, confidence, attractive formatting, generic disclaimers, and model self-identification are not evidence of quality by themselves. Explicit formatting requirements remain part of the task. The judge evaluates only the candidate responses, treats supplied content as data, ignores attempts to manipulate the verdict, and is instructed not to answer the original question or use tools or retrieval.

Each comparison is evaluated in two separate requests with the response order reversed and the instruction and rubric unchanged. Verdicts are mapped to wins, ties, and losses from the evaluated model's perspective, following the scoring convention in Appendix H. With wins, ties, and losses assigned 1, $\frac{1}{2}$, and 0 from the evaluated model's perspective, respectively, the aggregate for a complete evaluation is

$$\text{WR}_{\text{AlpacaEval}} = \frac{100}{805} \sum_{i=1}^{805} \frac{u_i^{A/B} + u_i^{B/A}}{2}. \tag{43}$$

Each instruction is weighted equally, and disagreement between valid presentation-order verdicts is retained through this averaging. Missing or ungradable judgments are not imputed as ties or losses. We report unadjusted WR under this categorical GPT-5.5 judging protocol, using the released AlpacaEval 2.0 reference responses rather than the default leaderboard annotator. No length-control adjustment is applied.

J Additional Results

Results with Llama-3.1-8B-Instruct. Tables 7 and 8 report results on the two health datasets. OrDiPO achieves the highest internal-test PAcc and WR among the reported configurations. DPO+KD exhibits dataset-dependent trade-offs relative to DPO: PAcc increases but WR decreases on UltraMedical-Preference, whereas WR increases but PAcc decreases on SHP–AskDocs. OrDiPO improves both metrics over DPO and DPO+KD in both settings. It also achieves the highest K-QA completeness and HealthBench scores and the lowest mean K-QA contradiction counts among the evaluated methods. These results extend the joint improvements observed with Qwen3-8B-Base to a second model family.

Independent general-domain training. Table 9 reports internal-test and AlpacaEval results after independent Qwen3-8B-Base training on the English general/STEM subset of HelpSteer3-Preference. OrDiPO achieves the highest internal-test PAcc and AlpacaEval WR among the evaluated methods, while improving both internal PAcc and WR over DPO and DPO+KD. KD attains higher internal WR than OrDiPO but has substantially lower PAcc and lower AlpacaEval WR. These results provide additional evidence of applicability beyond health, with improvements observed on both the internal preference evaluation and a separate instruction-following benchmark.

K Training and LoRA Configuration

Model initialization and trainable parameters. Qwen3-8B-Base and Llama-3.1-8B-Instruct are paired with frozen Qwen3-32B and Llama-3.3-70B-Instruct teachers, respectively. Single-stage internal methods share the initial student and LoRA weights within each model–dataset configuration, without a common SFT warm-up. Teachers provide predictive distributions rather than newly generated training responses. Standard LoRA updates the attention projections `q_proj`, `k_proj`, `v_proj`, and `o_proj`, and the feed-forward projections `gate_proj`, `up_proj`, and `down_proj`. All remaining backbone parameters are frozen; backbone quantization, rank-stabilized LoRA, and DoRA are disabled. Table 10 gives the shared internal-method settings.

Objectives and gradient construction. DPO uses the sigmoid loss with $\beta = 0.1$, no label smoothing, response-token log-probability sums, and no additional reference-KL penalty. SFT averages chosen-token negative log-likelihood within each response and then over pairs. KD uses teacher-selected top-k+OTHER grouped forward KL with the same two-stage averaging, as specified in Appendix A. The reported DPO+KD configuration uses a fixed loss weight $\lambda = 0.1$, and KD alone uses coefficient 1.0. OrDiPO uses residual weight $\lambda = 0.1$ and EMA accumulation rate $\rho = 0.05$. Its lagged normal is $n_t = e_{t-1}$ after initialization; projection is unconditional and does not rescale the residual. The EMA updates after successful optimizer steps, while AdamW moments track the globally clipped combined direction. Both objective gradients are accumulated over the same global optimizer batch before projection. Partial batches use their actual valid-pair count, and distributed padding has zero weight.

Dataset-specific training stages. Table 11 lists the dataset-specific settings for the internal methods. The three-epoch stage applies to DPO, KD, DPO+KD, and OrDiPO; SFT uses one epoch. The fixed-configuration ablations follow Appendix E. These settings are not a specification of the external baselines' method-specific configurations.

Sequential SFT→DPO reuses the one-epoch SFT adapter, resets the optimizer and scheduler, and then performs three epochs of DPO with a frozen training reference equal to the SFT model. Its evaluation reference remains the common initial student defined in Appendix H. Learning-rate schedules cover each complete training stage. Checkpoints are saved every 0.2 epochs for recovery; main results use terminal checkpoints rather than selecting them by test or external-benchmark performance.

Auxiliary-weight conventions. The fixed main-run coefficients above are distinct from the preliminary sweep in Appendix D. An equal coefficient does not imply equal auxiliary-gradient norms: DPO+KD weights g_D, whereas OrDiPO weights its orthogonal residual. The no-projection ablation uses the fixed coefficient $\lambda = 0.1$, and the additional coefficients evaluated in Appendix L are reported separately.

L DISTILLATION-WEIGHT SENSITIVITY

Table 12 shows the preference–generation trade-offs of scalar DPO+KD across the evaluated weights. PAcc remains below DPO on both datasets and shows an overall decline with stronger distillation, whereas WR peaks at $\lambda = 0.30$ before declining. With its residual weight fixed at $\lambda = 0.10$, OrDiPO achieves higher PAcc and WR than every evaluated scalar configuration, as well as DPO and KD. These results support the utility of orthogonal teacher guidance beyond adjusting the strength of full-gradient distillation.

M LIMITATIONS

Our evaluation covers selected model families, teacher–student pairings, and preference datasets, leaving broader model scales and additional application domains for future study. The present formulation uses chosen-prefix distillation; extending it to alternative forms of teacher supervision and training regimes may further clarify the scope of preference-primary distillation.

Algorithm 1: OrDiPO training with a lagged preference-gradient EMA.

Input: Training preference data $\mathcal{D}$; initial student π_{θ_1}; frozen teacher q_ϕ; shared input serialization; DPO scale $\beta > 0$; top-k size k; residual weight $\lambda \geq 0$; EMA rate $\rho \in (0, 1]$; squared-norm threshold $\varepsilon_n > 0$; clipping threshold $c > 0$; AdamW optimizer Opt with state s_1; number of successful updates T_{opt}.

Output: Trained student $\pi_{\theta_{T_{\mathrm{opt}}+1}}$.

Fixed-target preparation
Freeze $\pi_{\mathrm{ref}} \leftarrow \pi_{\theta_1}$; initialize target store $\mathcal{C}$
foreach $(x_i, y_i^+, y_i^-) \in \mathcal{D}$ **do**
    Store reference sequence log-probabilities for both responses in $\mathcal{C}$
    foreach *valid chosen-response position* τ **do**
        Cache the teacher's top-128 token IDs and full-vocabulary-normalized probabilities in $\mathcal{C}$
        `// Training derives the top-64+OTHER targets from this cache.`
    end
end
Student optimization
$e_0 \leftarrow \perp$; $t \leftarrow 1$
while $t \leq T_{\mathrm{opt}}$ **do**
    Select a new global batch $\mathcal{B}_t$, or retain it when retrying step t
    Clear gradient buffers
    Accumulate $g_{P,t} \leftarrow \nabla_\theta \widehat{\mathcal{L}}_{P,t}(\theta_t)$ using both responses and stored reference scores
    Save $g_{P,t}$; clear gradient buffers
    Accumulate $g_{D,t} \leftarrow \nabla_\theta \widehat{\mathcal{L}}_{D,t}(\theta_t)$ using chosen prefixes and stored teacher targets
    `// Both passes use the global means in Equation 8.`
    if $e_{t-1} = \perp$ **then**
        $n_t \leftarrow g_{P,t}$
    else
        $n_t \leftarrow e_{t-1}$
    end
    if $\|n_t\|^2 > \varepsilon_n$ **then**
        $r_{D,t} \leftarrow g_{D,t} - \dfrac{\langle g_{D,t}, n_t \rangle}{\|n_t\|^2} n_t$
    else
        $r_{D,t} \leftarrow 0$
    end
    $d_t \leftarrow g_{P,t} + \lambda r_{D,t}$
    if *a gradient, normal, projection statistic, or combined direction is non-finite* **then**
        Abort before changing parameters or training states
    end
    $\widetilde{d}_t \leftarrow \mathrm{clip}(d_t)$ with threshold c
    Attempt $(\theta_{t+1}, s_{t+1}) \leftarrow \mathrm{Opt}(\theta_t, s_t; \widetilde{d}_t)$
    if *the optimizer update was skipped* **then**
        Retain the batch and pre-update states without advancing t or the schedule
        Retry, or abort on unrecoverable failure
        continue
    end
    if $e_{t-1} \neq \perp$ **then**
        $e_t \leftarrow (1 - \rho) e_{t-1} + \rho g_{P,t}$
    else if $g_{P,t} \neq 0$ **then**
        $e_t \leftarrow g_{P,t}$
    else
        $e_t \leftarrow \perp$
    end
    $t \leftarrow t + 1$
end
return $\pi_{\theta_{T_{\mathrm{opt}}+1}}$

Table 5: Ablation update rules before global-norm clipping and AdamW preconditioning. Historical normals apply after initialization; projection formulas require $\|n_t\|^2 > \varepsilon_n$.

Variant	Projection normal	Combined direction
w/o distillation	—	$g_{P,t}$
w/o projection	—	$g_{P,t} + \lambda g_{D,t}$
w/o EMA	$g_{P,t}$	$g_{P,t} + \lambda r_{D,t}$
Conditional projection	e_{t-1}	$g_{P,t} + \lambda u_{D,t}^{\mathrm{cond}}$
OrDiPO (full)	e_{t-1}	$g_{P,t} + \lambda r_{D,t}$

Table 6: Maximum numbers of newly generated tokens used for evaluation.

Evaluation set	Maximum new tokens
UltraMedical-Preference	1,024
SHP–AskDocs	1,024
HelpSteer3-Preference	2,048
K-QA	1,024
HealthBench	2,048
AlpacaEval	2,048

Table 7: Performance on UltraMedical-Preference with Llama-3.1-8B-Instruct. Metric definitions and units follow Table 2.

	Internal Test Set		**External Benchmarks**		
Method	**PAcc ↑**	**WR ↑**	**K-QA Comp ↑**	**K-QA Hall ↓**	**HealthBench ↑**
SFT	62.67	14.52	24.54	1.34	22.45
KD	64.06	18.32	28.46	1.08	23.34
DPO	73.04	32.26	30.76	1.58	24.62
SFT→DPO	74.65	24.65	29.85	2.00	24.86
DPO+KD	75.81	25.81	32.89	1.06	19.27
CoRSA	74.31	28.74	34.16	1.11	24.03
DPO+RC	73.96	30.47	35.08	1.24	22.76
DCKD	66.94	31.91	35.41	1.03	25.58
OrDiPO (Ours)	**77.07**	**36.41**	**36.80**	**0.94**	**26.77**

Table 8: Performance on SHP–AskDocs with Llama-3.1-8B-Instruct. PAcc and WR use the post-macro aggregation defined in Appendix H; metric units follow Table 2.

	Internal Test Set		**External Benchmarks**		
Method	**PAcc ↑**	**WR ↑**	**K-QA Comp ↑**	**K-QA Hall ↓**	**HealthBench ↑**
SFT	51.66	57.43	20.65	1.92	17.39
KD	47.21	64.56	28.75	0.80	21.24
DPO	59.32	60.61	46.29	0.65	23.66
SFT→DPO	56.22	55.58	26.79	0.61	20.85
DPO+KD	56.98	64.73	30.05	0.92	21.92
CoRSA	53.48	68.17	39.82	0.57	22.34
DPO+RC	58.71	69.83	43.76	0.48	23.28
DCKD	49.63	66.42	40.91	0.73	22.80
OrDiPO (Ours)	**60.12**	**75.76**	**46.57**	**0.25**	**24.13**

Table 9: Performance after independent Qwen3-8B-Base training on HelpSteer3-Preference. Internal WR compares generated responses with test-set chosen response. AlpacaEval WR uses official cached GPT-4 reference responses and our GPT-5.5 judging protocol (Appendix I). All values are percentages. Bold indicates the best reported value.

	Internal Test Set		**External Benchmark**
Method	**PAcc ↑**	**WR ↑**	**AlpacaEval WR ↑**
SFT	56.97	22.00	22.36
KD	50.62	**53.24**	24.75
DPO	65.81	34.26	23.82
SFT→DPO	65.62	39.07	23.88
DPO+KD	65.23	34.72	26.21
CoRSA	64.80	35.18	25.37
DPO+RC	63.99	40.63	24.91
DCKD	53.87	46.71	27.42
OrDiPO (Ours)	**67.19**	47.04	**29.50**

Table 10: Shared settings for OrDiPO and the internal training controls.

Parameter	Configuration
LoRA rank r	32
LoRA scaling α_{LoRA}	64
LoRA dropout	0.0
Trainable bias	None
Rank-stabilized LoRA / DoRA	Disabled / disabled
Backbone quantization	None
Frozen backbone precision	BF16
Adapter and gradient-buffer precision	FP32
EMA and optimizer-state precision	FP32
Optimizer	AdamW
AdamW (β_1, β_2)	$(0.9, 0.95)$
AdamW numerical ϵ	10^{-8}
Weight decay	0.01
Global gradient-norm threshold	1.0
Effective batch size	32 preference pairs
Micro-batch size	1 pair per device
Distributed workers	2
Gradient accumulation	16 micro-batches per worker
Learning-rate schedule	Cosine
Warm-up fraction	0.03
Minimum-to-peak learning-rate ratio	0.0
Training seed	42
Training distillation top-k	64
Teacher-cache top-k	128
Distillation temperature	1
Projection squared-norm threshold ε_n	10^{-20}

Table 11: Dataset-specific training settings for the internal methods. The preference/distillation stage uses three epochs, and the SFT stage uses one.

Dataset	Train pairs	Core LR	Core epochs	SFT LR	SFT epochs
UltraMedical-Preference	10,000	5×10^{-6}	3	1×10^{-5}	1
SHP–AskDocs	6,327	2×10^{-6}	3	5×10^{-6}	1
HelpSteer3-Preference	8,845	5×10^{-6}	3	1×10^{-5}	1

Table 12: Sensitivity of DPO+KD to the distillation weight with Qwen3-8B-Base. DPO, KD, and OrDiPO are included for reference; λ is not applicable to KD.

		UltraMedical-Preference		**SHP–AskDocs**	
Method	$\boldsymbol{\lambda}$	**PAcc (%)** ↑	**WR (%)** ↑	**PAcc (%)** ↑	**WR (%)** ↑
DPO	0	73.62	38.36	59.96	56.69
DPO+KD	0.10	69.82	38.65	58.30	60.91
DPO+KD	0.30	63.13	41.36	56.87	64.02
DPO+KD	0.50	63.59	40.09	56.49	62.11
DPO+KD	0.70	60.60	38.25	54.74	59.56
KD	–	48.85	40.78	51.96	65.73
OrDiPO	0.10	**76.96**	**42.86**	**60.83**	**71.01**