Stabilizing Hard-Negative Preference Optimization for Medical LLMs

Anonymous ACL submission

Abstract

Hard-negative preference optimization should improve support for ground-truth preferred responses while limiting drift from the reference model. In Japanese medical instruction tuning, we find that standard DPO and IPO improve the preferred–dispreferred gap but can leave preferred-response likelihood below the reference model and induce growing sequence-level KL divergence. Motivated by this failure-mode analysis, we propose a stabilized IPO objective that combines preferred-response anchoring with explicit KL control. Ablations show that positive anchoring improves preferred-response support, while KL control suppresses reference-model drift; their combination preserves anchoring gains while further reducing drift. These results suggest that reliable medical preference optimization benefits from both preferred-response anchoring and drift control.

1 Introduction

In medical response generation, alignment must improve factual reliability without eroding useful behavior acquired during supervised fine-tuning (SFT), because factual errors, unsafe recommendations, or excessive deviation from the SFT model can make responses clinically unreliable (Singhal et al., 2023, 2025).

Direct Preference Optimization (DPO) and Identity Preference Optimization (IPO) are representative lightweight alternatives to the conventional reinforcement learning from human feedback (RLHF) pipeline (Ouyang et al., 2022; Rafailov et al., 2023; Gheshlaghi Azar et al., 2024). In medical instruction data, a natural way to target remaining hard failures after SFT is to construct preference pairs from prompts on which the SFT model still produces insufficient responses, using the ground-truth SFT response as the preferred response and the model-generated insufficient response as the dispreferred response. We refer to this setting as *hard-negative preference optimization*: preference optimization that targets difficult SFT samples on which an SFT-trained model still fails, while preserving SFT-acquired behavior. However, both the effectiveness and learning stability of this setting remain unclear.

The instability of pairwise preference objectives is related to a broader line of work showing that DPO-style objectives can suffer from increased out-of-distribution (OOD) actions (Xu et al., 2024) and insufficient improvement of preferred-response likelihood (Pang et al., 2024; Adler et al., 2024; Liu et al., 2024). Existing stabilization strategies include adding a negative log-likelihood (NLL) loss on preferred responses, penalizing preferred responses that fall below the initial model as in DPO-Positive (DPOP) (Pal et al., 2024), and using finer-grained or token-level regularization (Cho et al., 2025). Nevertheless, the learning dynamics of pairwise objectives in hard-negative preference optimization remain insufficiently characterized: they may improve the preferred–dispreferred gap while still failing to anchor the preferred response relative to the reference model or to prevent unnecessary drift away from the reference model.

As shown in **Fig. 1**, we analyze DPO/IPO training dynamics and develop IPOP+KL, a stabilized IPO objective combining preferred-response anchoring with explicit KL control. The two terms address complementary failures: under-anchoring of the ground-truth preferred response and reference-model drift not controlled by the preferred–dispreferred gap.

We evaluate IPOP+KL in a 13B Japanese medical SFT setting, with ablations, LLM-as-a-judge evaluation, a four-physician evaluation, and two auxiliary medical SFT baselines. While our study focuses on medical instruction-following, it highlights a broader challenge for knowledge-intensive alignment: pairwise preference optimization must improve preference gaps while preserving reli-

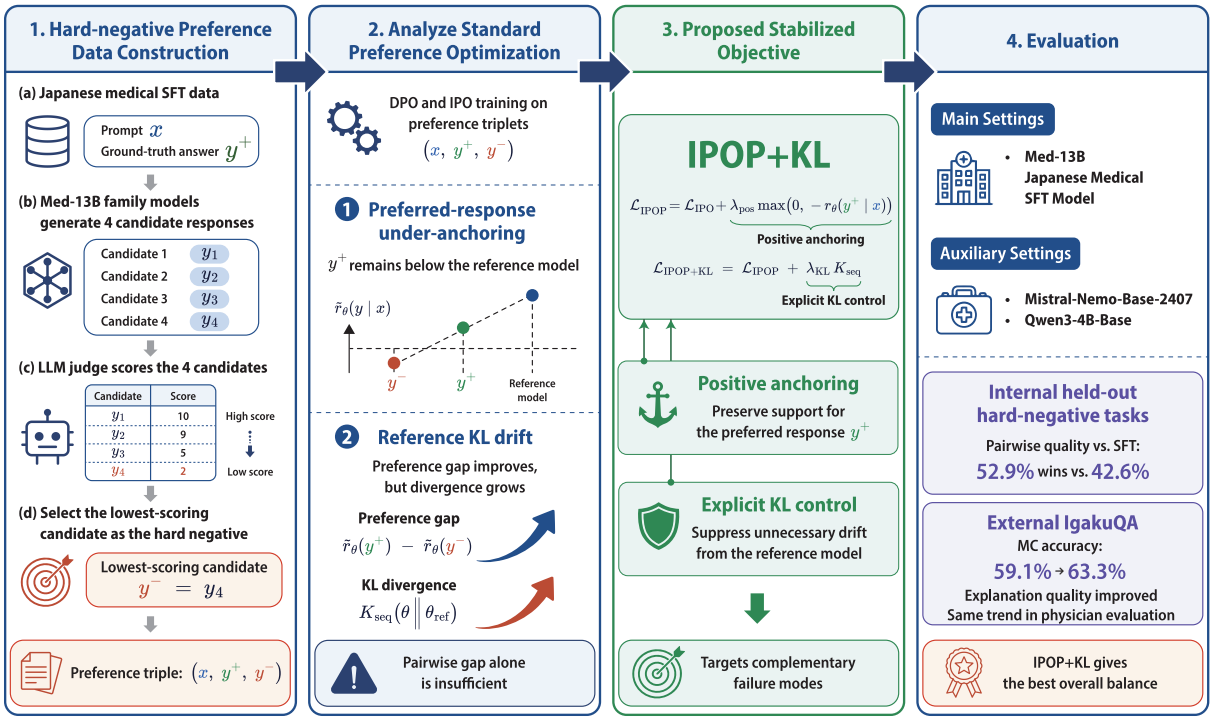

Figure 1: Study overview. We construct hard-negative preference triples from Japanese medical SFT data, analyze DPO/IPO failure modes, and propose IPOP+KL, which combines preferred-response anchoring with explicit KL control. We evaluate it on internal hard-negative medical tasks and external IgakuQA in medical LLM settings.

able preferred responses and limiting unnecessary reference-model drift.

Our contributions are threefold:

- We identify preferred-response under-anchoring and reference-KL drift as DPO/IPO failure modes in hard-negative medical instruction tuning.
- We show that positive anchoring and KL control provide complementary stabilization.
- We improve response quality on hard medical samples in a 13B Japanese medical SFT setting, with robustness checks on two auxiliary baselines and support from LLM-as-a-judge and physician evaluations.

2 Related Work

LLM Alignment and Preference Optimization. LLM alignment commonly starts with SFT and then applies preference-based optimization. Classical RLHF trains a reward model and then optimizes the policy with reinforcement learning under reference-model regularization (Ziegler et al., 2020; Stiennon et al., 2020; Ouyang et al., 2022). DPO replaces this pipeline with a direct pairwise objective relative to a reference model (Rafailov et al., 2023), while IPO fits a finite preferred–dispreferred margin instead of continually enlarging the gap (Gheshlaghi Azar et al., 2024). Related methods such as ORPO, KTO, SimPO, and online preference optimization explore alternative formulations of preference-based alignment (Hong et al., 2024; Ethayarajh et al., 2024; Meng et al., 2024; Calandriello et al., 2024).

Instability and Stabilization of Pairwise Preference Optimization. Recent studies show that DPO-style objectives can suffer from distributional drift, increased mass on OOD responses, and insufficient improvement of preferred-response likelihood (Xu et al., 2024; Pal et al., 2024; Cho et al., 2025; Fisch et al., 2025). Existing stabilization strategies include adding supervised or preferred-response likelihood terms (Pang et al., 2024; Adler et al., 2024; Liu et al., 2024), modifying the pairwise loss to penalize preferred-below-reference behavior or more effectively constrain rejected responses (Pal et al., 2024; Cho et al., 2025; Zeng et al., 2024), and strengthening reference-policy or KL-based regularization (Tang et al., 2024; Zeng et al., 2024; Liu et al., 2025).

Factuality Alignment and Medical Response Correction. Factuality alignment is crucial in

medical response generation because a plausible answer can be unacceptable if it contains factual errors, unsafe recommendations, or missing clinical caveats (Singhal et al., 2023, 2025). Prior work improves medical or long-form factuality through clinically aligned assistants, sentence-level supervision, uncertainty-aware evaluation, atomic preferences, and token-level regularization (Zhang et al., 2025; Kawakami et al., 2025; Lyu et al., 2026; Huang and Chen, 2024; Xue et al., 2025; Yuan et al., 2025; Zeng et al., 2024; Zhou et al., 2025). Another line performs post-hoc correction or edit-based rewriting after generation (Peng et al., 2023; Madaan et al., 2023; Balachandran et al., 2022; Li and Flanigan, 2025; Ben Abacha et al., 2025; Malmi et al., 2019; Mallinson et al., 2020).

Positioning of This Work. Rather than proposing a new regularization primitive, we analyze how existing stabilization principles behave in hard-negative medical preference optimization and combine preferred-response anchoring with KL control to address the identified failure modes.

3 Method

3.1 Problem Setting

Given a medical prompt x from the SFT training corpus and its ground-truth response y^+, we define a *hard-negative sample* as one for which an SFT-trained model still produces an insufficient response y^- after SFT. From each hard-negative sample, we construct a preference pair (x, y^+, y^-), using the ground-truth SFT response as preferred and the insufficient model-generated response as dispreferred. Our goal is to separate y^+ from y^- while preserving support for y^+ relative to the reference policy π_0 and limiting unnecessary reference-model drift.

3.2 Relative-Gap Formulation of Preference Optimization

Let π_0 denote the reference policy and π_θ the trainable policy. For a response $y = (y_1, \ldots, y_T)$ with token mask m_t, define

$$\begin{aligned} r_\theta(y \mid x) &= \tfrac{1}{Z_y} \textstyle\sum_{t=1}^{T} m_t \, \ell_t, \\ \ell_t &= \log \pi_\theta(y_t \mid x, y_{<t}) - \log \pi_0(y_t \mid x, y_{<t}). \end{aligned} \tag{1}$$

Here, $Z_y = \sum_t m_t$ when using average log-probabilities. The pairwise reward gap is

$$h_\theta(x, y^+, y^-) = r_\theta(y^+ \mid x) - r_\theta(y^- \mid x). \tag{2}$$

DPO and IPO both optimize the trainable policy through this relative gap. Following DPO, we write the pairwise loss as

$$\mathcal{L}_{\text{DPO}} = -\log \sigma\big(\beta \, h_\theta(x, y^+, y^-)\big), \tag{3}$$

where $\beta > 0$ is the reference-regularization parameter that scales the log-ratio gap.

Following IPO, we write the finite-margin loss as

$$\mathcal{L}_{\text{IPO}} = \left(h_\theta(x, y^+, y^-) - \frac{1}{2\tau} \right)^2, \tag{4}$$

where $\tau > 0$ determines the target margin $1/2\tau$. Thus, DPO keeps encouraging a larger preferred–dispreferred gap through the logistic loss, whereas IPO fits a finite target gap. Because both objectives depend only on $h_\theta(x, y^+, y^-)$, they do not directly constrain whether the preferred response is improved relative to the reference model, nor do they directly penalize drift that preserves this pairwise score.

3.3 Why Pairwise Objectives Can Miss Residual Drift

We next identify two limitations of relative-gap objectives. In the unnormalized sequence-probability view, the pairwise gap can be written as

$$\begin{aligned} h_\theta(x, y^+, y^-) = \log \tfrac{\pi_\theta(y^+|x)}{\pi_\theta(y^-|x)} \\ - \log \tfrac{\pi_0(y^+|x)}{\pi_0(y^-|x)}. \end{aligned} \tag{5}$$

This sequence-level view shows that DPO and IPO compare preferred and dispreferred responses by their likelihood ratios relative to the reference model, rather than by their absolute probabilities.

Limitation 1: lack of absolute anchoring. A favorable pairwise gap does not require the preferred response to become more likely than it was under the reference model. Equivalently, optimizing h_θ does not require

$$r_\theta(y^+ \mid x) \geq 0. \tag{6}$$

Thus, the model can improve the preference gap by decreasing the rejected response, even if the preferred response remains under-anchored relative to π_0.

Limitation 2: residual-mass null direction. The same relative-gap structure also leaves some residual redistributions unconstrained. As an illustrative sequence-level example, divide the response

space into three parts: the preferred response y^+, the dispreferred response y^-, and all other continuations. If both observed responses are scaled by the same factor,

$$\begin{aligned} \pi_s(y^+ \mid x) &= s\,\pi_\theta(y^+ \mid x), \\ \pi_s(y^- \mid x) &= s\,\pi_\theta(y^- \mid x), \end{aligned} \tag{7}$$

their ratio is unchanged, and therefore the pairwise gap is unchanged. Especially when $s < 1$, the removed probability mass can then be reassigned to other continuations that have low probability under the reference model, which can be viewed as an OOD-like residual shift: residual mass can increase and reference-model divergence can grow without changing the DPO or IPO loss. Although the residual space is not directly observed in a finite preference batch, this construction motivates adding an explicit policy/reference deviation penalty on the observed preference responses as a tractable drift-control proxy. **Appendix A.1** provides a detailed mathematical explanation.

3.4 Stabilized IPO Objective

Motivated by the structural limitations of relative-gap objectives, we build a stabilized objective on IPO, whose finite-margin formulation avoids the unbounded gap expansion encouraged by DPO. We add two complementary terms: absolute preferred-response anchoring and explicit sequence-level KL control. The former discourages preferred-below-reference solutions, while the latter penalizes drift not controlled by the preferred–dispreferred gap.

To anchor the preferred response, we introduce

$$\mathcal{L}_{\text{pos}} = \max\left(0, -r_\theta(y^+ \mid x)\right), \tag{8}$$

which is active when the preferred response is below the reference model in log-ratio reward, analogous to the positive anchoring idea in **DPOP** (Pal et al., 2024). We refer to IPO augmented with this positive term as **IPOP**:

$$\mathcal{L}_{\text{IPOP}} = \mathcal{L}_{\text{IPO}} + \lambda_{\text{pos}}\mathcal{L}_{\text{pos}}. \tag{9}$$

Here λ_{pos} controls the strength of preferred-response anchoring.

As a tractable proxy for drift control, we add a sequence-level forward-KL penalty over valid preferred and dispreferred response sequences. For each valid response token, define

$$\begin{aligned} \delta_t &= \log \pi_0(y_t \mid x, y_{<t}) - \log \pi_\theta(y_t \mid x, y_{<t}), \\ k_t &= \exp(\delta_t) - \delta_t - 1. \end{aligned} \tag{10}$$

Following the `k3` forward-KL estimator (Schulman, 2020), k_t serves as a nonnegative token-level policy/reference divergence proxy. For each response sequence y, we compute

$$K_{\text{seq}}(x, y) = \frac{1}{\sum_t m_t} \sum_t m_t k_t, \tag{11}$$

where m_t masks valid response tokens. Let $\mathcal{B}_{\text{seq}}$ be the set of preferred and dispreferred response sequences in a batch. The KL-control term is

$$\mathcal{L}_{\text{KL}} = \lambda_{\text{KL}} \frac{1}{|\mathcal{B}_{\text{seq}}|} \sum_{(x,y)\in\mathcal{B}_{\text{seq}}} K_{\text{seq}}(x, y), \tag{12}$$

where λ_{KL} controls the strength of KL regularization. This term penalizes policy/reference divergence on valid preference-response tokens. Additional details on token masking and sequence-level aggregation are provided in **Appendix A.2**.

The final stabilized objective, **IPOP+KL**, is

$$\mathcal{L}_{\text{IPOP+KL}} = \mathcal{L}_{\text{IPOP}} + \mathcal{L}_{\text{KL}}. \tag{13}$$

Equivalently, it combines finite-margin preference optimization, absolute preferred-response anchoring, and explicit reference-model regularization in a single objective.

We do not add an auxiliary NLL loss to this objective, although such losses have been used to stabilize DPO in prior work (Liu et al., 2024). This choice isolates the effect of the proposed preference-objective modifications.

4 Experiments

We evaluate IPOP+KL in a main 13B Japanese medical SFT setting and two auxiliary medical SFT settings based on `mistralai/Mistral-Nemo-Base-2407`[1] and `Qwen/Qwen3-4B-Base`[2]. The main setting is used for learning-dynamics analysis, ablations, and response-quality evaluation; auxiliary settings test response-quality robustness.

4.1 Main Experimental Setup

The main preference-learning experiment starts from a publicly released 13B-parameter Japanese medical reasoning LLM; we refer to it as `Med-13B` hereafter.[3] This model has already undergone

[1] https://huggingface.co/mistralai/Mistral-Nemo-Base-2407

[2] https://huggingface.co/Qwen/Qwen3-4B-Base

[3] The HuggingFace identifier of the model is omitted to preserve anonymity and will be restored in the camera-ready version.

SFT on a large-scale Japanese medical instruction-tuning corpus formatted with a thinking section followed by the final answer. The SFT corpus contains 1.44M examples, corresponding to 2.79B training tokens,[4] and covers Japanese medical QA, clinical reasoning, clinical-document processing, medical translation, summarization, and general instruction following. We provide the category-level SFT corpus composition in **Appendix B.1**. In our experiments, we use the SFT-completed `Med-13B` checkpoint as both the medical SFT baseline and the reference policy π_0 for preference optimization.

4.2 Hard-Negative Preference Data Construction

We construct preference pairs from the original SFT training data. For each sampled SFT example, we use the ground-truth SFT response as the preferred response y^+. To obtain candidate dispreferred responses, we generate responses with temperature $= 0.7$ from four variants in the same medical LLM family: the standard `Med-13B` model, a weaker `Med-13B` variant trained on a smaller SFT subset of roughly 0.4M examples and 1.1B tokens, a weaker `Med-8x13B` variant trained on the same SFT subset, and the standard `Med-8x13B` model trained on the full SFT corpus, an 8×13B mixture-of-experts variant from the same model family. Using both standard and weaker variants increases the diversity of generated errors while keeping the prompts and response format close to the target medical SFT distribution.

We score each generated response using `openai/gpt-oss-120b`[5], given the user query and the ground-truth SFT response. The judge assigns a 10-point score based on medical accuracy, contextual appropriateness, alignment with the ground-truth response, and instruction following. For each prompt, we select the generated response with the lowest score among the four medical LLM family candidates as the rejected response y^-; in the broad hard-negative analysis, generated responses with a score below 8 are treated as insufficient. The resulting preference triple is (x, y^+, y^-), where y^+ is the ground-truth SFT response and y^- is the insufficient model-generated response for a hard-negative sample. We use the ground-truth SFT response, rather than another sampled model response, as y^+ because even high-scoring model generations may contain subtle factual errors, missing caveats, or unsupported claims. **Appendix B.2** details the preference-data construction process and score-threshold analysis.

4.3 Baselines, Comparisons, and Ablations

We compare the proposed objective with continued-SFT and standard preference-optimization baselines. The continued-SFT baseline (cSFT) trains only on the preferred ground-truth SFT responses and does not use rejected responses. We also evaluate cSFT+KL, which adds the same sequence-level forward-KL penalty against the original SFT reference policy, to test whether reference regularization alone is sufficient without pairwise preference learning. For preference optimization, we compare DPO, DPOP, IPO, IPO+KL, IPOP, and IPOP+KL. We report detailed hyperparameters, including learning rate, batch size, and distributed training settings, in **Appendix C.1**.

4.4 Auxiliary Base-Model Experiments

To test whether the response-quality trend extends beyond `Med-13B`, we conduct two auxiliary experiments using `Mistral-Nemo-Base-2407` and `Qwen3-4B-Base`. Each base model is first instruction-tuned for two epochs on 300,000 examples randomly subsampled from the same Japanese medical SFT corpus, and the resulting model is used as both the initialization and reference policy for preference learning. For tractability, we reuse the preference-pair set mined with `Med-13B` rather than re-mining hard negatives for each auxiliary model. For each auxiliary baseline, we evaluate cSFT, DPO, IPO, IPOP, and IPOP+KL using the same preference-learning duration and objective-level hyperparameters as in the main experiment where applicable; auxiliary SFT and preference-learning settings are reported in **Appendix C.2** and **Appendix C.3**, respectively.

4.5 Evaluation Protocol

We evaluate training behavior, internal response quality, IgakuQA generalization, and physician preferences.

4.5.1 Training Dynamics

We monitor pairwise preference accuracy, the reward gap h_θ, preferred and dispreferred reward means, and the sequence-level reference-KL proxy. Detailed metric definitions are provided in **Appendix D.1**.

[4] Throughout this paper, token counts are computed using the tokenizer of `Med-13B`.

[5] https://huggingface.co/openai/gpt-oss-120b

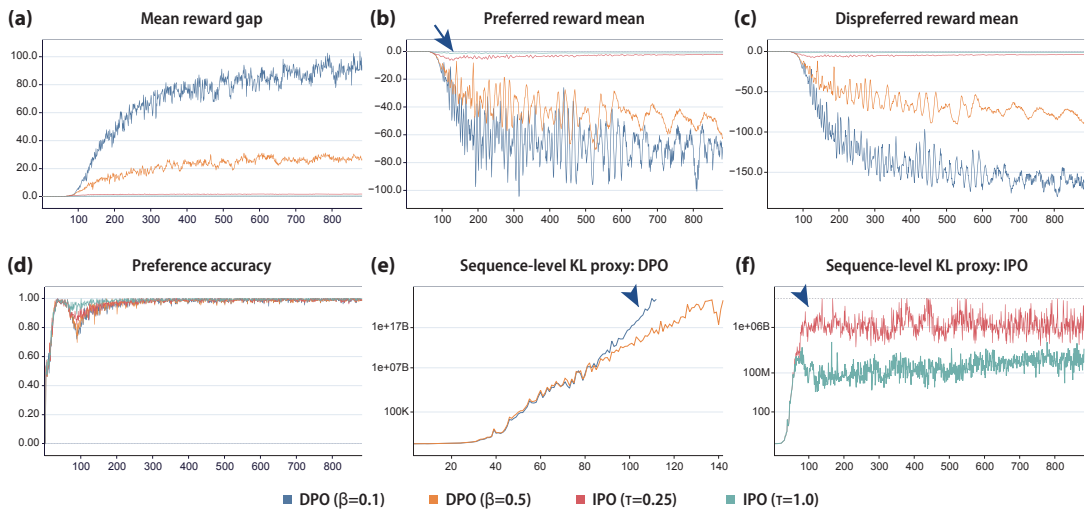

Figure 2: Training dynamics of standard DPO and IPO objectives. DPO and IPO improve the pairwise preference signal but leave preferred-response rewards below the reference level (see the arrow in **(b)**) and induce substantial reference-KL drift (see the arrowheads in **(e)** and **(f)**). Extreme KL spikes are clipped for visualization.

4.5.2 Internal Response-Quality Evaluation

We evaluate generated-response quality on 8,592 held-out hard-negative examples reserved from the hard-negative candidate pool and not used for preference-learning training. We conduct pairwise comparisons and 0–10 per-response scoring using `Qwen/Qwen3.5-397B-A17B-FP8` as the judge. Dataset composition, judge prompts, and inference settings are provided in **Appendix D.2**–**Appendix D.4**.

4.5.3 External Response-Quality Evaluation

We use the IgakuQA benchmark (Kasai et al., 2023), a Japanese medical licensing examination dataset that is not included in the preference-learning dataset, to evaluate external generalization. We report multiple-choice answer accuracy and pairwise free-form explanation-quality comparisons against the SFT baseline and among top methods. As a supplementary sensitivity analysis, we also repeat the main IgakuQA pairwise comparisons with `google/gemini-2.5-flash`. Additional details are provided in **Appendix D.5**.

4.5.4 Physician Evaluation

Four physicians with 6–16 years of clinical experience compared anonymized, randomly ordered IgakuQA responses from the reference model and IPOP+KL. We used 20 shared questions to compute Fleiss' κ and 30 non-overlapping questions per physician to analyze aggregate preferences over 120 judgments.

5 Results

5.1 Standard Pairwise Objectives Do Not Ensure Preferred-Response Anchoring or Control Drift

We first examine preferred-response under-anchoring and reference-model drift under standard pairwise objectives. **Fig. 2** shows DPO/IPO training dynamics in the main `Med-13B` setting.

DPO and IPO successfully optimize the pairwise signal: preference accuracy rapidly saturates, and the preferred–dispreferred reward gap increases. However, the preferred reward remains negative throughout training, indicating that the preferred response stays below the reference model (see the arrow in **Fig. 2b**). The reward decomposition shows that the gap is driven largely by decreasing dispreferred rewards rather than by increasing preferred rewards.

The sequence-level reference-KL proxy further increases or exhibits large spikes, showing that standard pairwise objectives do not prevent drift from the reference model (see the arrowheads in **Fig. 2ef**). Thus, standard DPO and IPO improve relative preference gaps but do not ensure preferred-response anchoring or drift control, motivating the stabilized objective introduced in **Section 3.4**.

5.2 Positive Anchoring and KL Control Address Complementary Failure Modes

We next ablate the two stabilization components motivated by the failure modes identified

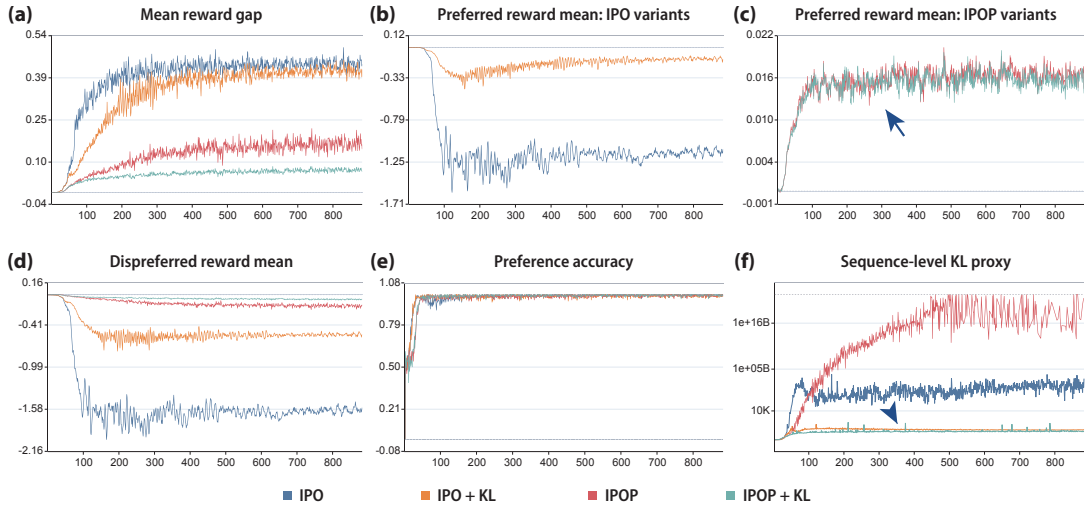

Figure 3: Ablation of preferred-response anchoring and KL control. Positive anchoring improves preferred-response support (see the arrow in **(c)**), while KL control suppresses reference-KL drift (see the arrowhead in **(f)**); IPOP+KL addresses both failure modes. Extreme KL spikes are clipped for visualization.

above. **Fig. 3** compares IPO, IPO+KL, IPOP, and IPOP+KL in the main `Med-13B` setting.

IPO improves the preferred–dispreferred reward gap and quickly reaches high preference accuracy, but its preferred-response reward remains far below the reference model. IPO+KL suppresses the sequence-level reference-KL proxy (see the arrowhead in **Fig. 3f**), yet still fails to anchor the preferred response. Thus, KL control alone limits drift but does not solve preferred-response under-anchoring.

Adding the positive anchoring term changes this behavior: IPOP and IPOP+KL keep preferred-response rewards near or above the reference level (see the arrow in **Fig. 3c**). However, IPOP alone still permits substantial growth in the sequence-level reference-KL proxy, indicating that anchoring alone does not control distributional drift.

IPOP+KL combines the strengths of both terms, preserving preferred-response support while keeping reference-KL drift substantially lower than IPOP. These results show that positive anchoring and KL control address complementary failure modes. Thus, adding KL control to IPOP primarily improves drift control while preserving the preferred-response anchoring achieved by the positive term. We provide an additional direct comparison between DPO and DPOP in **Appendix E.1**.

5.3 Stabilized IPO Improves Internal Response Quality

We evaluate generated-response quality on the held-out internal hard-negative medical task dataset. **Table 1** reports pairwise comparisons against the reference model, i.e., the SFT-completed model used as π_0 in the main `Med-13B` setting.

In the main internal pairwise comparison, KL control improves response quality most when the underlying objective exhibits substantial reference-model drift. Adding KL to cSFT leaves quality essentially unchanged (49.68% for cSFT vs. 48.62% for cSFT+KL), whereas adding KL to standard IPO raises the method win rate from 33.70% to 50.65%. This suggests that KL control mainly recovers quality by preventing drift-induced degradation. Positive anchoring provides a comparable recovery, with IPOP reaching 52.81%, and combining anchoring with KL yields the highest win rate among fully evaluated methods, with IPOP+KL reaching 52.98%. The small gain over IPOP is consistent with the same interpretation: once anchoring has reduced drift-induced degradation, less quality loss remains for KL to recover. Head-to-head internal comparisons in **Appendix E.2** further show that IPOP+KL is marginally favored over IPOP and favored over DPOP.

Per-response scoring in **Appendix E.3** also shows the same tendency: IPOP+KL achieves the highest overall score among fully evaluated methods, and adding KL to IPOP improves all five scoring axes in the main `Med-13B` setting, with the

Method	Reference wins	Method wins	Tie
cSFT	46.13	49.68	4.19
cSFT+KL	46.61	48.62	4.77
DPO ($\beta = 0.5$)*	97.25	2.34	0.41
DPO ($\beta = 0.1$)*	91.99	7.62	0.39
DPOP	45.08	51.24	3.68
IPO ($\tau = 1.0$)	65.14	33.70	1.16
IPO ($\tau = 0.25$)	84.97	14.62	0.41
IPO+KL	45.67	50.65	3.67
IPOP	43.83	52.81	3.36
IPOP+KL	42.69	**52.98**	4.32

Table 1: Internal pairwise response-quality evaluation in the main `Med-13B` setting. Each method is compared against the reference model; values are win/tie rates among parsed judgments. *DPO was evaluated only partially because degenerate outputs, including meaningless repetition, caused early termination.

Comparison	Judge	Left	Right	Tie
Ref. vs. DPOP	Primary	36.85	**43.90**	19.25
	Gemini	32.52	**44.72**	22.76
Ref. vs. IPOP	Primary	37.32	**44.32**	18.36
	Gemini	32.82	**46.12**	21.06
Ref. vs. IPOP+KL	Primary	37.55	**43.10**	19.35
	Gemini	32.27	**44.57**	23.16
DPOP vs. IPOP	Primary	38.29	**40.44**	21.27
	Gemini	35.70	**38.25**	26.05
DPOP vs. IPOP+KL	Primary	38.82	**41.67**	19.51
	Gemini	36.00	**38.85**	25.15
IPOP vs. IPOP+KL	Primary	**41.67**	37.87	20.46
	Gemini	36.85	**37.65**	25.50

Table 2: IgakuQA free-form explanation-quality comparison by the primary and alternative judges. Left/Right refer to the comparison column; presentation order was randomized and hidden.

largest gains on medical accuracy and safety.

In auxiliary experiments, IPOP and IPOP+KL show close, metric-dependent results, with internal head-to-head comparisons slightly favoring IPOP and scoring/IgakuQA results tending to favor IPOP+KL. Detailed auxiliary results are provided in **Appendix E.4** and **Appendix E.5**.

5.4 Stabilized IPO Generalizes to IgakuQA

We next evaluate whether the internal response-quality gains generalize to IgakuQA, an external Japanese medical licensing-examination benchmark. In multiple-choice accuracy, IPOP+KL improves the main `Med-13B` reference model from 59.10% to 63.30%, slightly outperforming cSFT+KL, IPOP, and DPOP. Consistent with the internal response-quality pattern, adding KL control to standard IPO raises this more objective metric from 46.80% to 60.70%, while the gain from IPOP to IPOP+KL is smaller but positive. The same trend holds in the auxiliary `Mistral-Nemo` and `Qwen3-4B` settings, where IPOP+KL also achieves the highest accuracy within each base-model block. Detailed accuracy results are provided in **Appendix E.6**.

We also evaluate free-form explanation quality on IgakuQA using pairwise LLM-as-a-judge comparisons. **Table 2** reports the main `Med-13B` comparisons that were evaluated by both the primary judge and Gemini. IPOP+KL, IPOP, and DPOP are all preferred over the reference model by both judges, showing a similar tendency under the supplementary alternative-judge check. Among these strong methods, IPOP+KL remains broadly competitive with IPOP and is favored over DPOP by both judges. Full primary-judge IgakuQA pairwise results, including auxiliary base-model experiments, are provided in **Appendix E.7**.

A four-physician evaluation showed fair but low agreement on 20 shared questions (Fleiss' $\kappa = 0.207$) and a preference distribution of 28.3% reference wins, 41.7% IPOP+KL wins, and 30.0% ties over 120 non-overlapping judgments. Excluding ties, IPOP+KL showed a favorable but not significant trend over the reference model (59.5% vs. 40.5%; $p = 0.051$, one-sided binomial test), consistent with the LLM-as-a-judge results.

6 Conclusions

We identified preferred-response under-anchoring and reference-model drift as key failure modes of standard DPO/IPO in hard-negative medical preference optimization, and developed IPOP+KL, which combines preferred-response anchoring and explicit KL control to address them. Across internal and IgakuQA evaluations, preferred-response anchoring accounts for much of the response-quality recovery, while KL control contributes mainly by preventing drift-induced degradation. This quality effect is largest for standard IPO, where drift is substantial, and smaller for IPOP, where positive anchoring already mitigates much of the drift-induced loss. Even in this anchored regime, adding KL control improves all per-response quality axes in the main `Med-13B` setting, with the largest gains on medical accuracy and safety. Overall, reliable medical preference optimization should improve preferred responses while preserving reference-grounded behavior and limiting unnecessary drift, especially in safety-sensitive medical generation where unsupported, incomplete, overconfident, or unsafe explanations can be clinically problematic.

Limitations

First, IPOP+KL does not uniformly outperform IPOP across settings. Although it shows the strongest overall trend, IPOP remains competitive in some comparisons, suggesting that optimal reference-model regularization may depend on the reference model and evaluation setting.

Second, our response-quality evaluation primarily relies on LLM-as-a-judge protocols. Larger-scale physician evaluation is needed to validate the observed improvements for clinically nuanced long-form answers.

Finally, our experiments focus on Japanese medical instruction following, and we do not claim full cross-lingual or cross-domain generality. Future work should test the same failure-mode analysis and objective design in English medical QA, multilingual clinical QA, and non-medical knowledge-intensive domains.

Ethical Considerations

This study develops research models for medical instruction following and is not intended for direct clinical or patient-facing deployment. The study did not involve patients or newly collected patient data. A limited physician evaluation was conducted on anonymized model outputs for research assessment; no patient data or individually identifiable medical information was collected. All training and evaluation data were derived from publicly available medical resources, public benchmarks, open-access literature, drug information, and synthetic data, and to the best of our knowledge contain no individually identifiable patient information. Accordingly, the study was considered outside the scope of institutional ethical review.

References

Bo Adler, Niket Agarwal, Ashwath Aithal, Dong H. Anh, Pallab Bhattacharya, Annika Brundyn, Jared Casper, Bryan Catanzaro, Sharon Clay, Jonathan Cohen, Sirshak Das, Ayush Dattagupta, Olivier Delalleau, Leon Derczynski, Yi Dong, Daniel Egert, Ellie Evans, Aleksander Ficek, Denys Fridman, and 62 others. 2024. Nemotron-4 340B Technical Report. *arXiv preprint arXiv:2406.11704*.

Vidhisha Balachandran, Hannaneh Hajishirzi, William Cohen, and Yulia Tsvetkov. 2022. Correcting Diverse Factual Errors in Abstractive Summarization via Post-Editing and Language Model Infilling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9818–9830.

Asma Ben Abacha, Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Meliha Yetisgen, Fei Xia, and Thomas Lin. 2025. MEDEC: A Benchmark for Medical Error Detection and Correction in Clinical Notes. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22539–22550.

Daniele Calandriello, Zhaohan Daniel Guo, Remi Munos, Mark Rowland, Yunhao Tang, Bernardo Avila Pires, Pierre Harvey Richemond, Charline Le Lan, Michal Valko, Tianqi Liu, Rishabh Joshi, Zeyu Zheng, and Bilal Piot. 2024. Human alignment of large language models through online preference optimisation. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24.

Jae Hyeon Cho, JunHyeok Oh, Myunsoo Kim, and Byung-Jun Lee. 2025. Rethinking DPO: The Role of Rejected Responses in Preference Misalignment. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 8159–8176.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FLASHATTENTION: fast and memory-efficient exact attention with IO-awareness. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Model alignment as prospect theoretic optimization. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24.

Adam Fisch, Jacob Eisenstein, Vicky Zayats, Alekh Agarwal, Ahmad Beirami, Chirag Nagpal, Peter Shaw, and Jonathan Berant. 2025. Robust Preference Optimization through Reward Model Distillation. *Transactions on Machine Learning Research*, pages 1–27.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A General Theoretical Paradigm to Understand Learning from Human Preferences. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 4447–4455.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. ORPO: Monolithic Preference Optimization without Reference Model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189.

Chao-Wei Huang and Yun-Nung Chen. 2024. FactAlign: Long-form Factuality Alignment of Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16363–16375.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. *Applied Sciences*, 11(14).

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. 2023. Evaluating GPT-4 and ChatGPT on Japanese Medical Licensing Examinations. *arXiv preprint arXiv:2303.18027*.

Wataru Kawakami, Keita Suzuki, and Junichiro Iwasawa. 2025. Stabilizing Reasoning in Medical LLMs with Continued Pretraining and Reasoning Preference Optimization. *arXiv preprint arXiv:2504.18080*.

Changmao Li and Jeffrey Flanigan. 2025. RAC: Efficient LLM Factuality Correction with Retrieval Augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 25145–25159.

Yixin Liu, Pengfei Liu, and Arman Cohan. 2025. Understanding Reference Policies in Direct Preference Optimization. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8037–8052.

Zhihan Liu, Miao Lu, Shenao Zhang, Boyi Liu, Hongyi Guo, Yingxiang Yang, Jose Blanchet, and Zhaoran Wang. 2024. Provably mitigating overoptimization in RLHF: your SFT loss is implicitly an adversarial regularizer. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24.

Shiwei Lyu, Xidong Wang, Lei Liu, Hao Zhu, Chaohe Zhang, Jian Wang, Jinjie Gu, Benyou Wang, and Yue Shen. 2026. ClinAlign: Scaling Healthcare Alignment from Clinician Preference. *arXiv preprint arXiv:2602.09653*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. SELF-REFINE: iterative refinement with self-feedback. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23.

Jonathan Mallinson, Aliaksei Severyn, Eric Malmi, and Guillermo Garrido. 2020. FELIX: Flexible Text Editing Through Tagging and Insertion. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1244–1255.

Eric Malmi, Sebastian Krause, Sascha Rothe, Daniil Mirylenka, and Aliaksei Severyn. 2019. Encode, Tag, Realize: High-Precision Text Editing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5054–5065.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. SimPO: simple preference optimization with a reference-free reward. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260.

Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. 2024. Smaug: Fixing Failure Modes of Preference Optimisation with DPO-Positive. *arXiv preprint arXiv:2402.13228*.

Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. 2024. Iterative reasoning preference optimization. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. PyTorch: an imperative style, high-performance deep learning library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback. *arXiv preprint arXiv:2302.12813*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: your language

model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23.

John Schulman. 2020. Approximating KL divergence. Blog post.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and 13 others. 2023. Large language models encode clinical knowledge. *Nature*, 620:172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, and 16 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31:943–950.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20.

Yunhao Tang, Daniel Zhaohan Guo, Zeyu Zheng, Daniele Calandriello, Rémi Munos, Mark Rowland, Pierre Harvey Richemond, Michal Valko, Bernardo Ávila Pires, and Bilal Piot. 2024. Generalized preference optimization: a unified approach to offline alignment. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24.

Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. 2024. Is DPO superior to PPO for LLM alignment? a comprehensive study. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24.

Boyang Xue, Fei Mi, Qi Zhu, Hongru Wang, Rui Wang, Sheng Wang, Erxin Yu, Xuming Hu, and Kam-Fai Wong. 2025. UAlign: Leveraging Uncertainty Estimations for Factuality Alignment on Large Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6002–6024.

Hongbang Yuan, Yubo Chen, Pengfei Cao, Zhuoran Jin, and Kang Liu. 2025. Beyond Under-Alignment: Atomic Preference Enhanced Factuality Tuning for Large Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6325–6338.

Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. 2024. Token-level direct preference optimization. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24.

Yiming Zhang, Zheng Chang, Wentao Cai, MengXing Ren, Kang Yuan, Yining Sun, and Zenghui Ding. 2025. IIMedGPT: Promoting Large Language Model Capabilities of Medical Tasks by Efficient Human Preference Alignment. *arXiv preprint arXiv:2501.02869*.

Wenxuan Zhou, Shujian Zhang, Lingxiao Zhao, and Tao Meng. 2025. T-REG: Preference Optimization with Token-Level Reward Regularization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27876–27889.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-Tuning Language Models from Human Preferences. *arXiv preprint arXiv:1909.08593*.

A Methodological Details

A.1 Residual-Mass Null Direction

This appendix provides a mathematical explanation of the residual-mass null direction discussed in **Section 3.3**. Consider a response space consisting of the preferred response, the dispreferred response, and a residual set of all other continuations. Let $\mathcal{O}(x)$ denote the residual output space not covered by the observed pair:

$$\mathcal{O}(x) = \mathcal{Y} \setminus \{y^+, y^-\}. \tag{14}$$

Define the corresponding residual mass under the trainable policy as

$$M_\theta^{\text{res}}(x) = \sum_{y \in \mathcal{O}(x)} \pi_\theta(y \mid x). \tag{15}$$

Consider a simplified output space in which

$$\pi_\theta(y^+ \mid x) = a, \qquad \pi_\theta(y^- \mid x) = b, \tag{16}$$

and hence

$$M_\theta^{\text{res}}(x) = 1 - a - b. \tag{17}$$

For any $s \in (0, 1)$, define a new policy π_s by shrinking the probabilities of both observed responses by the same factor:

$$\pi_s(y^+ \mid x) = sa, \qquad \pi_s(y^- \mid x) = sb, \tag{18}$$

and redistributing the removed mass $(1 - s)(a + b)$ to $\mathcal{O}(x)$. Then

$$h_s(x, y^+, y^-) = \log \frac{sa}{sb} - \log \frac{\pi_0(y^+ \mid x)}{\pi_0(y^- \mid x)}. \tag{19}$$

Since $sa/sb = a/b$, we have $h_s(x, y^+, y^-) = h_\theta(x, y^+, y^-)$. Hence the pairwise DPO and IPO losses are unchanged along this redistribution direction:

$$\mathcal{L}_{\mathrm{IPO}}(x; \pi_s) = \mathcal{L}_{\mathrm{IPO}}(x; \pi_\theta), \tag{20}$$

and

$$\mathcal{L}_{\mathrm{DPO}}(x; \pi_s) = \mathcal{L}_{\mathrm{DPO}}(x; \pi_\theta). \tag{21}$$

At the same time, the residual mass becomes

$$M_s^{\mathrm{res}}(x) = 1 - s(a + b), \tag{22}$$

which is larger than $M_\theta^{\mathrm{res}}(x)$ whenever $s < 1$. This is an idealized construction, but it captures the key invariance of pairwise objectives: preserving the preferred–dispreferred ratio is sufficient to preserve the pairwise loss, regardless of how the remaining probability mass is allocated.

We can also make the KL effect explicit. Let $a_0 = \pi_0(y^+ \mid x)$ and $b_0 = \pi_0(y^- \mid x)$. Under the redistribution above,

$$R_s(x) = \sum_{y \in \mathcal{O}(x)} \pi_s(y \mid x) \log \frac{\pi_s(y \mid x)}{\pi_0(y \mid x)}, \tag{23}$$

so that

$$\begin{aligned} &D_{\mathrm{KL}}(\pi_s(\cdot \mid x) \,\|\, \pi_0(\cdot \mid x)) \\ &\quad = sa \log \tfrac{sa}{a_0} + sb \log \tfrac{sb}{b_0} + R_s(x). \end{aligned} \tag{24}$$

This quantity is not generally invariant to s. If the redistributed mass is assigned to continuations in $\mathcal{O}(x)$ for which $\pi_0(y \mid x)$ is small, then the residual term can become large. Thus residual mass can increase while the pairwise DPO/IPO loss remains unchanged, and the reference KL can increase as well. We refer to this as OOD-like drift when the moved mass enters low-reference-probability regions of the residual space.

This construction does not imply that optimization must follow this exact redistribution path. Rather, it shows that the pairwise objective itself provides no penalty against such redistributions.

A.2 Additional Details on KL Control

Section 3.4 defines the KL-control term used in IPOP+KL. Here, we provide additional implementation details on token masking, sequence-level aggregation, and the interpretation of the KL proxy.

For each preferred or dispreferred response sequence, the token mask m_t excludes prompt tokens and padding tokens, so the KL proxy is computed only over valid response tokens. When a response has no valid response tokens after masking, it is excluded from the sequence-level aggregation. In practice, valid sequences are collected into $\mathcal{B}_{\mathrm{seq}}$, which contains both preferred and dispreferred responses from valid preference pairs.

The sequence-level proxy $K_{\mathrm{seq}}(x, y)$ is computed by averaging token-level k_t values over valid response tokens within each sequence. The batch-level KL-control term then averages $K_{\mathrm{seq}}(x, y)$ uniformly over sequences in $\mathcal{B}_{\mathrm{seq}}$, rather than weighting sequences by their lengths. Thus, each valid preferred or dispreferred response contributes equally to the batch-level KL penalty.

The quantity $k_t = \exp(\delta_t) - \delta_t - 1$ follows the `k3` forward-KL estimator and is nonnegative in exact arithmetic. In standard on-policy settings, this estimator is used with samples from the policy. In our setting, the tokens are teacher-forced preference-response tokens rather than fresh policy rollouts. We therefore use k_t as a stable token-level policy/reference divergence proxy on the observed preference responses, rather than as an unbiased rollout-based estimate of the full policy KL or a direct measurement of residual OOD continuations.

B Data Details

B.1 SFT Corpus Composition

This appendix summarizes the SFT corpus used to train `Med-13B` before preference learning. **Table B.1** presents the task-category mixture; examples denote training records, and token counts are aggregated over the same categories. All SFT samples are formatted with a thinking section followed by the final answer. The non-medical SFT data in the General instruction / dialogue and General reasoning / coding categories use publicly available Japanese and English instruction-tuning corpora.

General instruction / dialogue. This category covers broad instruction following, response formatting, safety-oriented refusal behavior, and multi-turn dialogue. It also includes synthetic chat-distillation data from strong instruction-tuned models to improve general conversational robustness.

Medical exam QA. This category contains Japanese and English medical QA benchmark training samples and exam-style multiple-choice data. Public or commonly used benchmark training sources include MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), and PubMedQA (Jin et al.,

Category	Examples	Tokens	Description
General instruction / dialogue	311,926	547.7M	Instruction following, formatting, safety behavior, greetings, and multi-turn dialogue.
Medical exam QA	436,028	291.1M	Training samples from medical QA benchmarks such as MedQA, MedMCQA, and PubMedQA, plus past Japanese National Medical Practitioners Qualifying Examination questions that do not overlap with IgakuQA.
Clinical extraction / conversion	52,043	106.8M	Structured conversion of clinical text and extraction of drugs, labs, background, treatments, and other attributes.
Clinical reasoning / assessment	36,510	85.8M	Differential diagnosis, laboratory abnormality interpretation, prognosis, problem lists, and follow-up planning.
Medical long-form QA	345,962	955.2M	Long-answer QA derived from guidelines, drug information, PMC Open Access articles, retrieval-grounded synthetic QA, and other medical documents.
Translation / bilingual biomedical text	114,102	524.0M	English–Japanese medical translation and biomedical literature translation.
Summarization / patient-facing generation	31,637	61.3M	Patient-facing documents, simplified medical documents, and short or structured summaries.
General reasoning / coding	116,048	218.0M	Auxiliary mathematical, logical, coding, and broad instruction-following data.
Total	1,444,256	2.79B	

Table B.1: Category-level composition of the Japanese medical SFT corpus. By example count, the three largest categories are medical exam QA, medical long-form QA, and general instruction/dialogue; by token count, medical long-form QA is the largest category.

2019). The mixture also includes Japanese medical, nursing, and pharmacy examination questions, including past Japanese National Medical Practitioners Qualifying Examination questions that do not overlap with IgakuQA (Kasai et al., 2023).

Clinical extraction / conversion. This category includes tasks that convert clinical text into structured formats and extract medically relevant fields. These tasks include converting medical records into structured JSON-like outputs, producing Subjective, Objective, Assessment, and Plan (SOAP)-style representations, and extracting drugs, laboratory values, patient background, treatments, and other clinical attributes from free text.

Clinical reasoning / assessment. This category targets clinical reasoning rather than direct fact recall. It includes differential diagnosis, laboratory abnormality interpretation, patient risk and prognosis assessment, problem-list construction, and clinical follow-up plan generation. These sources expose the model to stepwise medical decision support and long-form clinical explanation patterns.

Medical long-form QA. This category contains knowledge-intensive long-answer QA derived from Japanese clinical guidelines, adult rare-disease guidelines, medical payment-system information, drug information, PubMed abstracts, PMC Open Access biomedical articles,[6] textbooks, and other medical documents. It also includes synthetic QA grounded in relevant medical documents, with particular attention to Japanese medical contexts. The retrieval-grounded sources are score-filtered to retain synthetic question–answer pairs with higher estimated quality and factuality.

Translation / bilingual biomedical text. This category includes both general medical translation and biomedical literature translation tasks. These tasks cover English-to-Japanese and Japanese-to-English translation, including biomedical article and medical-domain translation data.

Summarization / patient-facing generation. This category covers patient-facing document generation, simplified medical document generation, short-paragraph summarization, and structured summarization. These tasks complement QA and extraction by training the model to produce accessible explanatory text and concise clinical summaries.

General reasoning / coding. To preserve broad reasoning ability, this category includes non-medical or weakly medical general-reasoning sources such as logical, mathematical, and coding instruction data, including Japanese–English synthetic coding sources. These data are included as auxiliary instruction-following and reasoning supervision rather than as primary medical-domain evidence.

[6] https://pmc.ncbi.nlm.nih.gov/tools/openftlist/

Judge score range	Candidates
0–3	146,138
4–6	121,532
7–< 8	33,116
8–10	159,274

Table B.2: Score distribution for generated candidate responses during hard-negative mining. The record-level score is defined as the minimum judge score among the evaluated source-model generations for each prompt.

B.2 Preference Data Construction

Preference pairs are constructed as described in **Section 4.2**. Briefly, for each sampled SFT prompt, the ground-truth SFT response is used as the preferred response y^+, and the lowest-scoring medical LLM family generated response is used as the rejected response y^-. For the auxiliary base-model experiments, which serve as response-quality robustness checks, we reuse the same preference-pair set for computational tractability rather than remining hard negatives separately for each auxiliary baseline.

Scoring is performed with `openai/gpt-oss-120b` using deterministic inference and the prompt shown in **Appendix B.3**.

We use score < 8 as a broad hard-negative threshold, rather than restricting rejected responses to only severe failures. This threshold captures not only obvious errors but also clinically important omissions, subtle factual errors, unsupported claims, and missing caveats that may receive mid-range scores. The resulting score distribution is reported in **Table B.2**. Applying this threshold yields 300,786 candidate hard-negative samples; after excluding held-out examples for evaluation, 292,194 pairs are used as the preference-learning training dataset. The category-level composition of the training dataset is shown in **Table B.3**, and descriptive statistics for the resulting preference pairs are reported in **Table B.4**.

B.3 Hard-Negative Judge Prompt

The hard-negative mining judge uses `openai/gpt-oss-120b` with temperature 0.0. It receives the dialogue history, user query, reference answer, and generated candidate answer; assigns a 10-point score based on medical accuracy, contextual appropriateness, alignment with the reference answer, and instruction following; and returns a structured JSON output that can be parsed automatically. The English translation of the prompt template is shown below.

```
# Role
You are a top-tier expert in evaluating medical dialogue.
Analyze the provided information and strictly grade the
↪  quality of the target model.

{{history}}

<Target User Query>
{{question}}
</Target User Query>

<Correct Answer (Reference)>
{{reference_answer}}
</Correct Answer (Reference)>

<Model Generated Content (To be evaluated)>
{{generated_content}}
</Model Generated Content (To be evaluated)>

# Evaluation Criteria (10-point scale)
Grade on a 10-point scale, not a 100-point scale, according
↪  to the following criteria.

1. Medical Accuracy: [Most important]
    - If the answer contains statements that contradict
    ↪  facts, such as fictitious disease
      names, incorrect mechanisms, or treatments without
      ↪  evidence, apply a large penalty
      immediately (at least -5 points).
2. Contextual Appropriateness:
    - Does the answer follow the preceding conversation and
    ↪  read like a natural dialogue?
    - Does it follow meta-instructions such as "explain it
    ↪  conversationally"?
3. Alignment with Reference:
    - Does the answer cover the important points presented in
    ↪  the reference answer?
4. Instruction Following:
    - Does the answer follow the specified tone and format?

# Scoring Guide
- 10: Perfect. Medically accurate, with insight equal to or
↪  better than the reference,
      and extremely natural dialogue.
- 7-9: Generally good. No medical errors, but some
↪  information is missing relative to
        the reference, or the dialogue is slightly unnatural.
- 4-6: Important information is missing, or there are minor
↪  medical misunderstandings.
- 0-3: Serious hallucinations (false statements), or complete
↪  disregard of the dialogue context.

# Output Format (JSON)
Output the evaluation result only in the following JSON
↪  format. Include the reasoning process
in the analysis.
```json
{
  "Evaluation": "Provide detailed analysis based on each
  ↪  evaluation criterion. In particular, explicitly
  ↪  describe the comparison with the reference answer, the
  ↪  medical checks, and reasons for any deductions.",
  "Critical_Issues": ["Describe any issues if present;
  ↪  otherwise use an empty list"],
  "Score": 0.0,
  "Result": "Pass" or "Fail"
}
```
```

C Training Details

C.1 Main Experiment Hyperparameters

This appendix summarizes the training configuration used in the main `Med-13B` experiment. A compact summary of the training objectives compared in the main experiment is provided in **Table C.1**. Objective-specific hyperparameters are reported in **Table C.2**. All preference-optimization methods in

Category	Pairs	Preferred tokens	Avg. pref.	Dispreferred tokens	Avg. dispref.
General instruction / dialogue	54,309	48.2M	887.5	60.9M	1,121.6
Medical exam QA	64,769	48.2M	744.5	68.1M	1,050.9
Clinical extraction / conversion	24,278	37.2M	1,532.7	36.1M	1,486.0
Clinical reasoning / assessment	24,277	42.3M	1,740.9	40.7M	1,676.1
Medical long-form QA	75,975	96.5M	1,269.6	98.5M	1,295.8
Translation / bilingual biomedical text	16,756	46.7M	2,784.4	49.8M	2,969.4
Summarization / patient-facing generation	13,700	23.1M	1,684.9	18.7M	1,362.6
General reasoning / coding	18,130	25.8M	1,421.5	37.3M	2,057.5
Total	292,194	367.9M	1,258.9	409.9M	1,402.9

Table B.3: Category-level composition of the preference-learning dataset. The table reports the number of preference pairs, total preferred-response tokens, total dispreferred-response tokens, and average response length by category; Avg. pref. and Avg. dispref. denote the average preferred- and dispreferred-response lengths, respectively.

Metric	Mean / median
Token-level normalized edit distance	0.877 / 0.888
Character-level normalized edit distance	0.838 / 0.856
ROUGE-L F1	0.251 / 0.235
Length ratio dispreferred/preferred	1.275 / 1.008
Lexical overlap	0.254 / 0.242

Table B.4: Descriptive statistics for sampled preference pairs. Statistics are computed from a reservoir sample of 20,000 pairs drawn from the 292,194 available preference-learning training pairs.

Objective	Main hyperparameters
cSFT	continued SFT
cSFT+KL	$\lambda_{\mathrm{KL}} = 0.05$
DPO	$\beta \in \{0.5, 0.1\}$
DPOP	$\beta = 0.5$, $\lambda_{\mathrm{DPOP}} = 50$
IPO	$\tau \in \{1, 0.25\}$
IPO+KL	$\tau = 1$, $\lambda_{\mathrm{KL}} = 0.05$
IPOP	$\tau = 1$, $\lambda_{\mathrm{pos}} = 50$
IPOP+KL	$\tau = 1$, $\lambda_{\mathrm{pos}} = 50$, $\lambda_{\mathrm{KL}} = 0.05$

Table C.1: Training objectives compared in the main experiment. All methods are trained for one epoch from the same medical SFT baseline. For IPO, $\tau = 1$ and $\tau = 0.25$ correspond to target gaps of 0.5 and 2.0, respectively. Auxiliary base-model experiments evaluate cSFT, DPO, IPO, IPOP, and IPOP+KL as response-quality robustness checks.

the main experiment are trained for one epoch over the constructed preference dataset.

Software environment. All preference-learning runs use NVIDIA NeMo RL 0.5.0rc[7] with Python 3.12. The main runtime stack includes PyTorch 2.9.0 (Paszke et al., 2019), Transformers 4.57.1, Megatron-Core, Transformer Engine 2.8.0, FlashAttention 2.8.1 (Dao et al., 2022), and vLLM 0.11.2.

Model and tokenizer. The main experiment initializes the policy from `Med-13B`, which is also used as the reference policy for preference optimization. The tokenizer applies an instruction–response chat template with a fixed default system instruction and separate user and assistant headers.

Batching and sequence length. The global batch size is 128 preference triples. Because each triple contains one preferred and one dispreferred response, each global batch contains 256 response sequences. We use a micro-batch size of 1 and a maximum total sequence length of 32,768 tokens, with sequence lengths padded to multiples of 8.

Distributed training. Training is performed on 4 nodes with 8 NVIDIA H200 GPUs per node. Megatron-based distributed training uses tensor model parallel size 8, activation checkpointing, and no pipeline, context, expert-tensor, or expert-model parallelism.

Optimizer and scheduler. We use Adam with learning rate 8.0×10^{-7}, minimum learning rate 8.0×10^{-8}, weight decay 0.05, $\beta_1 = 0.9$, $\beta_2 = 0.95$, and $\epsilon = 10^{-8}$. The learning-rate schedule uses cosine decay with a warmup fraction of 0.05, and weight decay is held constant throughout training.

C.2 Auxiliary Base-Model SFT Hyperparameters

For the auxiliary experiments, we start from `mistralai/Mistral-Nemo-Base-2407` and `Qwen/Qwen3-4B-Base`. Each base model is first instruction-tuned for two epochs on 300,000 examples, corresponding to 20.8% of the medical SFT corpus, randomly subsampled from that corpus. Because neither model family is specifically tuned for Japanese medical long-form QA or for the response format assumed by our preference dataset, which includes a Japanese reasoning section before

[7] https://github.com/NVIDIA-NeMo/RL

the final answer, auxiliary SFT aligns both base models with the target domain and response format. The detailed auxiliary SFT setup is summarized in **Table C.3**. For each auxiliary model, we use the model-specific tokenizer and `chat_template`.

C.3 Auxiliary Preference-Learning Hyperparameters

The auxiliary medical SFT models are used as both initialization and reference policies for preference learning. For each auxiliary baseline, we evaluate cSFT, DPO, IPO, IPOP, and IPOP+KL. The preference-optimization runs use the same training duration and objective-level hyperparameters as in the main experiment where applicable. The auxiliary runs use the shared preference-pair set described in **Appendix B.2** and are evaluated as response-quality robustness checks.

D Evaluation Details

D.1 Monitoring Metrics

We monitor four diagnostic metrics to characterize pairwise preference fitting, preferred/dispreferred reward dynamics, and reference-model drift.

Preference accuracy. Preference accuracy is the fraction of valid preference pairs for which the preferred response has a larger log-ratio reward than the dispreferred response:

$$\mathrm{Acc}_{\mathrm{pref}} = \mathrm{mean}\left(\mathbf{1}[r_\theta(y^+ \mid x) > r_\theta(y^- \mid x)]\right). \tag{25}$$

This is a pairwise reward-gap metric and should not be interpreted as a generation-quality metric.

Mean reward gap. To measure the strength of pairwise separation, we report the mean preferred–dispreferred reward gap:

$$\mathrm{Gap} = \mathrm{mean}\left[r_\theta(y^+ \mid x) - r_\theta(y^- \mid x)\right]. \tag{26}$$

This metric captures how strongly the objective separates preferred responses from dispreferred responses.

Mean preferred and dispreferred rewards. We also report the mean log-ratio rewards for the preferred and dispreferred responses:

$$\begin{aligned} \mathrm{Reward}_{\mathrm{pref}} &= \mathrm{mean}\left[r_\theta(y^+ \mid x)\right], \\ \mathrm{Reward}_{\mathrm{dispref}} &= \mathrm{mean}\left[r_\theta(y^- \mid x)\right]. \end{aligned} \tag{27}$$

Together, these metrics decompose the reward gap and indicate whether preference optimization separates a pair by increasing support for the preferred response, decreasing support for the dispreferred response, or both.

Sequence-level KL mean. To monitor reference-model drift, we report the mean sequence-level KL proxy over valid response sequences:

$$\mathrm{KL}_{\mathrm{seq}} = \mathrm{mean}_{(x,y)}\left[K_{\mathrm{seq}}(x, y)\right]. \tag{28}$$

This metric summarizes policy/reference divergence on valid response tokens and is used as our main drift diagnostic.

D.2 Held-Out Internal Evaluation Dataset

We construct a held-out internal evaluation dataset by sampling from the held-out examples excluded during preference-data construction, as described in **Appendix B.2**, rather than from the preference-learning training set. The dataset contains 8,592 preference examples. **Table D.1** shows the category-level composition.

D.3 Internal Pairwise Comparison Protocol

For the internal response-quality evaluation described in **Section 4.5**, we first compare each preference-optimized model against its corresponding SFT baseline. For each base model, we then select the top three methods based on the baseline-comparison results and conduct all pairwise comparisons among the selected methods.

We use `Qwen/Qwen3.5-397B-A17B-FP8` as the pairwise judge under deterministic inference. The judge is served with tensor parallel size 8 and GPU memory utilization 0.85, using a maximum model length of 32,768 tokens. For each comparison, the judge is shown the user prompt and two anonymized responses. The assignment of model responses to the `A` and `B` positions is randomized and hidden from the judge, so the judge selects `A`, `B`, or `Tie` based only on response content.

The English translations of the Japanese prompts used for pairwise comparison are shown below.

System prompt.

```
You are a strict and fair medical QA evaluator.
Evaluate only the content of the answers, not the model names
↪  or presentation order.
Return only a JSON object.
```

User prompt template.

```
Compare the two responses to the following prompt and
↪  determine which one is better.

Evaluation criteria:
1. Medical accuracy
2. Directness in answering the question
```

Objective	β	τ	Target gap	λ_{pos}	λ_{KL}	NLL loss
cSFT	–	–	–	–	–	Yes
cSFT+KL	–	–	–	–	0.05	Yes
DPO	0.5 / 0.1	–	–	–	–	No
DPOP	0.5	–	–	50	–	No
IPO	–	1 / 0.25	0.5 / 2.0	–	–	No
IPO+KL	–	1	0.5	–	0.05	No
IPOP	–	1	0.5	50	–	No
IPOP+KL	–	1	0.5	50	0.05	No

Table C.2: Objective-specific hyperparameters. The main experiment evaluates all listed objectives, whereas the auxiliary base-model experiments evaluate cSFT, DPO, IPO, IPOP, and IPOP+KL. Continued SFT (cSFT) is treated as a separate baseline and is not mixed into the preference loss. NLL denotes negative log-likelihood loss.

Item	Value
Base models	`Mistral-Nemo-Base-2407` and `Qwen3-4B-Base`
SFT examples	300,000
Epochs	2
Fine-tuning type	Full-parameter
Learning rate	1.0×10^{-6} with cosine decay; minimum 1.0×10^{-7}; warmup fraction 0.05
Global batch size	128
Maximum sequence length	32,768 tokens
Optimizer	Adam; weight decay 0.01; $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-8}$; gradient clipping 1.0

Table C.3: Auxiliary SFT hyperparameters. All runs use one 8-GPU node, bfloat16 precision, sequence packing, tensor parallel size 4, and activation checkpointing.

Category	Pairs
General instruction / dialogue	1,192
Medical exam QA	1,300
Clinical extraction / conversion	800
Clinical reasoning / assessment	600
Medical long-form QA	1,500
Translation / bilingual biomedical text	600
Summarization / patient-facing generation	400
General reasoning / coding	200
Additional held-out hard-negative examples	2,000
Total	8,592

Table D.1: Category-level composition of the held-out internal evaluation dataset. The dataset is sampled separately from the preference-learning training set and is used for internal response-quality evaluation.

```
3. Logicality and consistency
4. Absence of unnecessary misinformation or serious errors
5. Clarity as Japanese text

Important rules:
- Choose "A" if Answer A is clearly better.
- Choose "B" if Answer B is clearly better.
- Choose "Tie" if there is no substantive difference in
↪  medical accuracy,
  comprehensiveness, or clarity, or if superiority cannot be
  ↪  determined.
- Evaluate only the content of the answers, not the model
↪  names or presentation order.
- The output must be a JSON object only.
- Do not output any explanatory text outside the JSON object.

Output format:
{
  "winner": "A", "B", or "Tie",
  "reason": "Briefly and specifically describe the reason for
  ↪  the judgment in Japanese"
}

[Question]
{{ question }}

[Answer A]
{{ answer_a }}

[Answer B]
{{ answer_b }}
```

D.4 Internal Per-Response Scoring Protocol

For per-response scoring, the judge independently evaluates each anonymized model response given the user prompt and reference response. It assigns 0–10 scores for medical accuracy, reference alignment, clinical utility, safety and calibration, format clarity, and overall quality; the overall-quality score is used as the primary summary. The judge model and inference settings are the same as those described in **Appendix D.3**. The English translations of the Japanese prompts used for per-response scoring are shown below.

System prompt.

```
You are a strict and fair medical QA evaluator.
Using the reference answer as the basis, score the target
↪  answer from a medical perspective.
Evaluate only the answer content, without being influenced by
↪  the target model name or generation source.
Return only a JSON object.
```

User prompt template.

```
Strictly evaluate the model answer to the following medical
↪  QA by comparing it with the reference answer.

Evaluation axes:
1. medical_accuracy:
   Medical accuracy. Are statements about disease concepts,
   ↪  pathophysiology,
   tests, diagnosis, treatment, prognosis, and risks
   ↪  accurate?
   Substantially penalize medical errors that contradict the
   ↪  reference answer,
   fictitious diseases or treatments, and unsupported
   ↪  assertions.

2. reference_alignment:
   Consistency with the reference answer. Does the answer
   ↪  appropriately cover
   the important medical points included in the reference
   ↪  answer?
```

```
   Penalize omissions of important points in the reference
   ↪  answer and
   explanations that contradict it.

3. completeness_and_clinical_utility:
   Completeness and clinical utility. Are important
   ↪  differentials, cautions,
   exceptions, and clinical caveats appropriately included?
   Do not assign a high score to an answer that is merely
   ↪  long but has low utility.

4. safety_and_calibration:
   Safety and handling of uncertainty. Does the answer avoid
   ↪  dangerous advice,
   overconfident claims, and wording that could mislead
   ↪  patients?
   When necessary, does it appropriately indicate the need
   ↪  for specialist
   consultation, additional testing, or individualized
   ↪  judgment?

5. format_and_clarity:
   Format adherence and clarity. Is the answer readable,
   ↪  logically organized,
   and clear as Japanese text?
   If a specific format is requested, does the answer follow
   ↪  it? Penalize
   verbosity, repetition, or broken formatting.

Scoring rules:
- Score each evaluation axis as an integer or decimal from 0
↪  to 10.
- Score overall_score from 0 to 10 as well.
- overall_score should not be a simple average; it should be
↪  an integrated
  assessment that especially emphasizes medical accuracy,
  ↪  consistency with
  the reference answer, and safety.
- If the answer contains a medically serious error,
↪  overall_score should in
  principle be 6 or lower, even if other axes are strong.
- If the answer presents a fictitious or nonexistent disease
↪  or treatment as
  fact, overall_score should in principle be 4 or lower.
- If the disease name or premise in the question is incorrect
↪  and the reference
  answer points this out, heavily penalize an answer that
  ↪  does not make that
  correction and instead responds definitively as if it were
  ↪  another disease.
- If the target answer includes <think>...</think>, evaluate
↪  both the reasoning
  process and the final answer.
- Do not output any explanatory text outside the JSON object.

Output format:
{
  "scores": {
    "medical_accuracy": 0.0,
    "reference_alignment": 0.0,
    "completeness_and_clinical_utility": 0.0,
    "safety_and_calibration": 0.0,
    "format_and_clarity": 0.0,
    "overall_score": 0.0
  },
  "critical_issues": [
    "Specifically describe any serious medical errors, safety
    ↪  issues, or broken formatting. Use an empty array if
    ↪  none are present."
  ],
  "reason": "Briefly and specifically describe the reason for
  ↪  the scores in Japanese"
}

[Question]
{{ question }}

[Reference Answer]
{{ reference_answer }}

[Target Answer]
{{ answer }}
```

D.5 External Response-Quality Evaluation Protocol

For IgakuQA, we report multiple-choice answer accuracy and pairwise explanation-quality judgments. The judge model and inference settings follow **Appendix D.3**. To compute multiple-choice answer accuracy, we use the same LLM to extract the answer option indicated in each target model's free-form response, and calculate accuracy as the fraction of examples for which the extracted option matches the gold option.

For free-form explanation quality, we use the same pairwise comparison protocol as in **Appendix D.3**. For each IgakuQA question, the judge compares two anonymized model responses, including the selected answer and accompanying explanation, with the A/B assignment randomized and hidden from the judge. Each model is first compared with its corresponding SFT baseline, and the top three methods are then compared in all pairs.

D.6 Supplementary Alternative-Judge Evaluation

For the supplementary alternative-judge evaluation on IgakuQA, we repeat the main pairwise comparisons using `google/gemini-2.5-flash`. We use deterministic inference with a temperature of 0.0 and a maximum output length of 16,384 tokens. The pairwise comparison prompt is identical to the one used in **Appendix D.3**.

E Additional Results and Examples

E.1 Additional DPO/DPOP Training Dynamics

Fig. E.1 provides additional training dynamics comparing DPO and DPOP. Consistent with the positive-anchoring motivation of DPOP, DPOP keeps the preferred-response reward close to the reference level, while DPO drives the preferred-response reward far below the reference model. This confirms that explicitly penalizing preferred-below-reference behavior can address preferred-response under-anchoring.

However, the sequence-level reference-KL proxy still grows substantially under DPOP. Thus, positive anchoring improves support for preferred responses (see the arrow in **Fig. E.1c**) but does not by itself prevent reference-model drift (see the arrowhead in **Fig. E.1f**). This result further supports the need to combine preferred-response anchoring with explicit KL control, as in IPOP+KL.

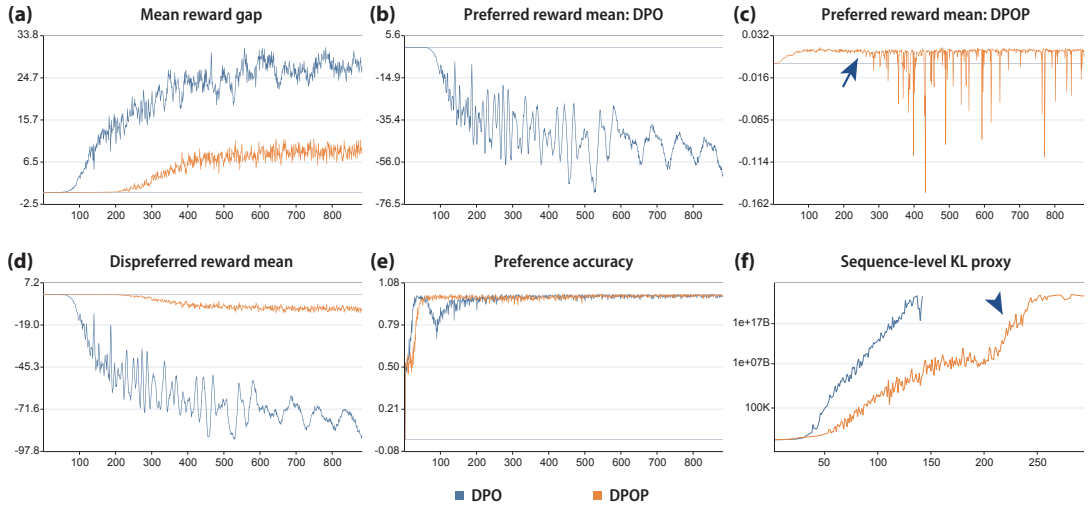

Figure E.1: Additional training dynamics for DPO and DPOP objectives. We compare DPO and DPOP in the main `Med-13B` setting. Preferred-response rewards are shown separately because their value ranges differ substantially. DPOP keeps preferred-response rewards near the reference level (see the arrow in **(c)**), whereas DPO drives them far below the reference model. However, DPOP still exhibits substantial growth in the sequence-level reference-KL proxy (see the arrowhead in **(f)**), indicating that positive anchoring alone does not control reference-model drift. Extreme spikes in the sequence-level KL proxy are clipped for visualization.

Comparison	Left wins	Right wins	Tie
DPOP vs. IPOP	45.74	**49.97**	4.29
DPOP vs. IPOP+KL	46.23	**49.81**	3.97
IPOP vs. IPOP+KL	47.91	**48.23**	3.86

Table E.1: Head-to-head internal pairwise comparisons among top main-setting methods. Left and Right refer to the model names as written in the comparison column; values are win/tie rates among parsed judgments.

E.2 Main-Setting Head-to-Head Internal Pairwise Evaluation

Table E.1 shows head-to-head comparisons among the strongest main-setting methods. IPOP+KL is competitive with IPOP and is favored over DPOP, while IPOP is also favored over DPOP. These results indicate that adding KL control to preferred-response anchoring does not reduce response quality in the main setting, while providing the more stable training dynamics shown in **Fig. 3**.

E.3 Main-Setting Internal Per-Response Scoring

Table E.2 reports per-response scoring results on the held-out internal hard-negative medical task dataset in the main `Med-13B` setting. The results are consistent with the pairwise comparison in **Table 1**: IPOP+KL obtains the highest overall score and the best scores across all reported axes. IPOP and IPO+KL also improve over the reference model, whereas DPO and IPO variants substantially degrade response quality.

E.4 Auxiliary Internal Pairwise Evaluation

Table E.3 reports auxiliary internal pairwise evaluations for the `Mistral-Nemo` and `Qwen3-4B` medical SFT baselines. Each trained model is compared against its corresponding reference model, namely the auxiliary SFT model used as π_0.

The auxiliary results are broadly consistent with the main setting. DPO substantially underperforms the reference model for both auxiliary baselines, and IPO also loses to the reference model. In contrast, IPOP and IPOP+KL are competitive with or improve over the reference model, indicating that preferred-response anchoring is important beyond the main `Med-13B` setting.

The auxiliary head-to-head comparisons in **Table E.4** show that IPOP is slightly favored over IPOP+KL in these auxiliary settings. Thus, while the auxiliary experiments support the importance of preferred-response anchoring for response quality, they also suggest that the best trade-off between anchoring and KL control may depend on the base model and training regime.

E.5 Auxiliary Internal Per-Response Scoring

Table E.5 reports per-response scoring results for the auxiliary `Mistral-Nemo` and `Qwen3-4B` settings. The results again show that DPO substan-

Method	Overall	Medical	Ref. align.	Format	Safety	Utility
Reference	6.395	6.501	5.880	8.711	7.263	6.688
cSFT	6.474	6.560	5.988	8.740	7.340	6.789
cSFT+KL	6.485	6.574	5.983	8.749	7.350	6.793
DPO ($\beta = 0.5$)*	0.037	0.050	0.019	0.000	0.057	0.002
DPO ($\beta = 0.1$)*	0.001	0.000	0.000	0.000	0.009	0.000
DPOP	6.429	6.513	6.020	8.623	7.295	6.795
IPO ($\tau = 1.0$)†	5.428	5.679	5.091	6.861	6.438	5.915
IPO ($\tau = 0.25$)	4.199	4.574	3.990	5.106	5.385	4.401
IPO+KL	6.490	6.557	6.049	8.728	7.363	6.819
IPOP	6.514	6.557	6.087	8.759	7.340	6.907
IPOP+KL	**6.595**	**6.659**	**6.151**	**8.801**	**7.438**	**6.964**

Table E.2: Internal per-response scoring in the main `Med-13B` setting. Scores are judged on a 0–10 scale; bold values indicate the best score in each column. *DPO was scored only partially due to degenerate outputs. †IPO with $\tau = 1.0$ was scored on a smaller subset.

Base model	Method	Reference wins	Method wins	Tie
Mistral-Nemo	cSFT	43.21	53.00	3.80
Mistral-Nemo	DPO	89.18	9.80	1.02
Mistral-Nemo	IPO	64.31	35.21	0.47
Mistral-Nemo	IPOP	42.63	**54.23**	3.13
Mistral-Nemo	IPOP+KL	43.26	53.03	3.72
Qwen3-4B	cSFT	45.46	47.84	6.70
Qwen3-4B	DPO	97.02	1.63	1.36
Qwen3-4B	IPO	75.38	23.81	0.81
Qwen3-4B	IPOP	43.19	**50.99**	5.81
Qwen3-4B	IPOP+KL	46.28	48.14	5.58

Table E.3: Auxiliary internal pairwise response-quality evaluation. Each method is compared against the corresponding reference model, i.e., the auxiliary medical SFT model used as π_0 for preference optimization. Values indicate win/tie rates among successfully parsed judgments. The auxiliary experiments serve as response-quality robustness checks rather than full replications of the main training-dynamics analysis.

tially degrades response quality, with near-zero overall scores. IPO also underperforms the reference model, particularly in format and overall quality.

In contrast, IPOP and IPOP+KL improve over the corresponding reference models in both auxiliary settings. For `Mistral-Nemo`, IPOP+KL obtains the highest overall score, while IPOP has slightly higher format quality. For `Qwen3-4B`, IPOP+KL obtains the highest overall score, with IPOP close behind. These scoring results support the auxiliary pairwise findings and suggest that preferred-response anchoring is consistently important for response quality, while the effect of KL control varies across base models.

E.6 IgakuQA Multiple-Choice Accuracy

Table E.6 reports multiple-choice accuracy on IgakuQA. Across all three base-model settings, IPOP+KL achieves the highest accuracy among the compared methods. DPO and IPO perform poorly, partly because their outputs often contain broken formats or missing extractable choices. These results support the external generalization trend discussed in **Section 5.4**.

E.7 IgakuQA Free-Form Pairwise Evaluation

Table E.7 reports IgakuQA free-form explanation-quality comparisons against the corresponding reference models. The results are consistent with the multiple-choice accuracy results in **Table E.6**. DPO and IPO substantially underperform the reference models, often due to degenerate, malformed, or incomplete outputs. In contrast, IPOP and IPOP+KL generally improve over the reference models across the main and auxiliary settings.

Table E.8 reports auxiliary head-to-head comparisons, where IPOP+KL is preferred over cSFT and slightly preferred over IPOP.

Base model	Comparison	Left wins	Right wins	Tie
Mistral-Nemo	cSFT vs. IPOP	47.49	**48.80**	3.72
Mistral-Nemo	cSFT vs. IPOP+KL	**48.88**	47.50	3.61
Mistral-Nemo	IPOP vs. IPOP+KL	**50.47**	46.08	3.45
Qwen3-4B	cSFT vs. IPOP	44.45	**49.56**	5.98
Qwen3-4B	cSFT vs. IPOP+KL	46.48	**47.95**	5.57
Qwen3-4B	IPOP vs. IPOP+KL	**49.36**	44.57	6.07

Table E.4: Auxiliary head-to-head internal pairwise comparisons. Left and Right refer to the model names as written in the comparison column; values are win/tie rates among parsed judgments.

Base model	Method	Overall	Medical	Ref. align.	Format	Safety	Utility
Mistral-Nemo	Reference	5.269	5.425	4.923	7.307	6.206	5.615
Mistral-Nemo	cSFT	5.406	5.433	4.962	**8.038**	6.276	5.753
Mistral-Nemo	DPO	0.042	0.051	0.048	0.012	0.172	0.032
Mistral-Nemo	IPO	4.051	4.378	3.974	5.091	5.207	4.439
Mistral-Nemo	IPOP	5.485	5.546	5.099	7.874	6.404	5.879
Mistral-Nemo	IPOP+KL	**5.514**	**5.624**	**5.164**	7.690	**6.459**	**5.903**
Qwen3-4B	Reference	5.084	4.925	4.647	8.247	5.820	5.401
Qwen3-4B	cSFT	5.135	4.966	4.683	**8.265**	5.889	5.427
Qwen3-4B	DPO	0.286	0.345	0.328	0.156	0.619	0.266
Qwen3-4B	IPO	3.349	4.228	3.820	2.831	4.458	3.642
Qwen3-4B	IPOP	5.120	5.104	4.888	7.751	5.986	5.482
Qwen3-4B	IPOP+KL	**5.263**	**5.135**	**4.957**	8.116	**6.117**	**5.654**

Table E.5: Auxiliary internal per-response scoring. Scores are judged on a 0–10 scale; bold values indicate the best score within each base-model block. IPOP+KL obtains the highest overall score for both auxiliary base models.

Base model	Method	Accuracy
Mistral-Nemo	Reference	52.80
Mistral-Nemo	cSFT	54.40
Mistral-Nemo	DPO*	0.05
Mistral-Nemo	IPO†	26.35
Mistral-Nemo	IPOP	52.95
Mistral-Nemo	IPOP+KL	**55.40**
Qwen3-4B	Reference	50.90
Qwen3-4B	cSFT	52.95
Qwen3-4B	DPO*	0.05
Qwen3-4B	IPO†	11.10
Qwen3-4B	IPOP	54.25
Qwen3-4B	IPOP+KL	**54.90**
Med-13B	Reference	59.10
Med-13B	cSFT	61.85
Med-13B	cSFT+KL	63.10
Med-13B	DPO ($\beta = 0.5$)*	0.05
Med-13B	DPO ($\beta = 0.1$)*	0.05
Med-13B	DPOP	62.15
Med-13B	IPO ($\tau = 1.0$)†	46.80
Med-13B	IPO ($\tau = 0.25$)†	31.75
Med-13B	IPO+KL	60.70
Med-13B	IPOP	62.05
Med-13B	IPOP+KL	**63.30**

Table E.6: IgakuQA multiple-choice accuracy. Accuracy is reported as a percentage over 2,000 questions. Bold values indicate the highest accuracy within each base-model block. *DPO and †IPO often produced degenerate or malformed outputs, resulting in many invalid or missing answer extractions.

Base model	Method	Reference wins	Method wins	Tie
Mistral-Nemo	cSFT	40.72	41.77	17.51
Mistral-Nemo	DPO*	94.68	2.86	2.46
Mistral-Nemo	IPO†	62.11	34.49	3.41
Mistral-Nemo	IPOP	40.37	43.57	16.06
Mistral-Nemo	IPOP+KL	38.79	**45.35**	15.87
Qwen3-4B	cSFT	40.99	44.29	14.71
Qwen3-4B	DPO*	96.17	2.31	1.52
Qwen3-4B	IPO†	68.87	28.14	2.99
Qwen3-4B	IPOP	40.67	**47.52**	11.81
Qwen3-4B	IPOP+KL	40.85	46.47	12.68
Med-13B	cSFT	36.52	41.92	21.56
Med-13B	cSFT+KL	37.47	42.87	19.66
Med-13B	DPO ($\beta = 0.5$)*	97.67	1.47	0.86
Med-13B	DPO ($\beta = 0.1$)*	88.69	7.60	3.70
Med-13B	DPOP	36.85	43.90	19.25
Med-13B	IPO ($\tau = 1.0$)†	62.99	31.90	5.11
Med-13B	IPO ($\tau = 0.25$)†	78.16	19.62	2.22
Med-13B	IPO+KL	39.52	42.97	17.51
Med-13B	IPOP	37.32	**44.32**	18.36
Med-13B	IPOP+KL	37.55	43.10	19.35

Table E.7: IgakuQA free-form explanation-quality comparisons against reference models. Values are win/tie rates among parsed judgments. *DPO and †IPO often produced degenerate, malformed, or incomplete outputs.

Base model	Comparison	Left wins	Right wins	Tie
Mistral-Nemo	cSFT vs. IPOP	40.99	**42.84**	16.17
Mistral-Nemo	cSFT vs. IPOP+KL	40.04	**42.74**	17.22
Mistral-Nemo	IPOP vs. IPOP+KL	41.35	**42.00**	16.65
Qwen3-4B	cSFT vs. IPOP	42.31	**44.21**	13.48
Qwen3-4B	cSFT vs. IPOP+KL	40.72	**45.07**	14.21
Qwen3-4B	IPOP vs. IPOP+KL	42.51	**44.47**	13.02

Table E.8: Auxiliary IgakuQA free-form head-to-head comparisons. Values are win/tie rates from the primary judge. Left and Right refer to the model names as written in the comparison column.