GAP-ADAPTIVE PER-OBJECTIVE REGULARIZATION AGAINST DIFFERENTIAL OVEROPTIMIZATION IN MULTI-OBJECTIVE ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-objective alignment scalarizes heterogeneous rewards into one objective optimized under a shared KL budget; as optimization strengthens, the policy overfits its reward proxies, and the resulting overoptimization is not distributed evenly across objectives. Existing methods either apply one global intervention (a single KL budget or cap) or only steer the direction of optimization, without adapting each objective's budget to a measured overoptimization gap. We formalize the answer in a minimal model of proxy and gold and name the pattern **Differential Overoptimization**: damage concentrates on the objective with the largest overoptimization susceptibility, and each objective's collapse point follows a closed-form allocation law. The train–held-out gap makes this damage observable during training—a bridge we derive in an extended model, up to one explicitly testable condition. Building on this, we propose **G**ap-**A**daptive **P**er-**O**bjective **R**egularization (**GAPOR**), one line per objective, with provable shrinkage, rebalancing, and closed-loop convergence that is global in the reward-model regime and bounded-gain in DPO. Simulations confirm the law, and experiments across multiple LLM families with independent gold evaluation show GAPOR removes the hackable objective's overoptimization gap while protecting true gold quality. We also analyze the shared-channel geometry our benchmark instantiates and derive a falsifiable collapse criterion; the sweep confirms its qualitative trajectory (plateau, turnover, and anti-correlation with the exact signal) and a fit of the gold sweep verifies the inequality quantitatively.

1 INTRODUCTION

Aligning a language model means optimizing several objectives at once—helpfulness, harmlessness, honesty, conciseness—each specified by its own reward signal (Yao et al., 2024). The standard approach, from weighted RLHF (Ouyang et al., 2022) to multi-objective DPO (Zhou et al., 2024), is to scalarize: sum the per-objective losses with the user's weights under one KL regularizer with one temperature β. A recent line foregoes scalarization: gradient-adaptive methods rebalance per-objective gradients each step to reach Pareto solutions matching user-specified trade-offs (Li et al., 2025).

Reward signals are imperfect proxies, so optimizing them hard raises the proxy while true gold quality falls—reward overoptimization, whose first-rises-then-falls gold curve is established for the single-objective case (Gao et al., 2022). Whether the objective is a scalarized sum or a conflict-balanced gradient combination, the signals remain proxies, and the multi-objective case raises the question this line of work leaves open: when K objectives share one budget, how does the overoptimization split among them?

We answer it in closed form and verify the observable consequences. Overoptimization does not split evenly: it concentrates on the objective with the largest overoptimization susceptibility (user weight $\times$ proxy spurious sensitivity $\times$ gold fragility), as the closed-form law of Section 2 makes precise. Exact signals are essentially immune within the orthogonal-channel model, though on a shared channel an exact signal can be dragged down by a conflicting soft proxy (Section 2.3); noisy, subjective, or length-confounded signals absorb most of the damage while their true quality degrades. We term this pattern differential overoptimization—a Goodhart effect specialized to the multi-objective template (Manheim & Garrabrant, 2018; Skalse et al., 2022). Viewing the scalarized

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.

template as a budget-allocation problem—each objective's share of the shared KL budget governs its overoptimization—yields a law, a bridge, and a one-line remedy.

Our contributions can be summarized as follows:

- **A law.** Modeling the scalarized template as a budget-allocation problem, we derive a closed-form allocation law—$\Delta_i \propto \gamma_i w_i b_i$—plus the collapse point $D_i^\star$, and a regime-corrected budget–temperature correspondence (Lemma A.9); three falsifiable predictions follow (hump shape, allocation ratio, collapse order). The law is derived in the β_i^{-1} reward-model regime, while DPO operates in the β_i^{-2} regime (Section 3).
- **A bridge.** The train–held-out margin gap tracks per-objective damage; we derive the bridge up to one explicitly testable condition (Condition 2.6, Theorem 2.7, Protocol P2).
- **A remedy.** The law identifies the per-objective budget as the place to act; scaling the DPO temperature by the observed gap, $\beta_i = \beta\,(1+\eta\,\hat{g}_i/s)$, gives GAPOR, one line per objective, with provable shrinkage and rebalancing, and closed-loop convergence that is global in the β_i^{-1} regime and bounded-gain in the β_i^{-2} regime (Appendix A).
- **A shared-channel theory.** Our benchmark instantiates a geometry the orthogonal-channel model excludes: helpfulness and the exact length rule are anti-correlated on one shared channel, where an exact signal caps the hackable channel or becomes the engine of its damage (Proposition 2.9). This yields a falsifiable collapse criterion $w_1 b_1/(w_2 b_2) > 1 + \tilde{b}_1/(2\gamma_1)$—whose qualitative trajectory the sweep confirms and whose inequality we verify by fitting the gold sweep (Section 4.3)—and a non-overshooting gain range (Remark A.7).

The full related-work discussion is in Appendix C.

2 Where Overoptimization Lands: An Allocation Law

2.1 Setup

The scalarized template. We study the scalarized template

$$\mathcal{L}(\theta) \;=\; \sum_{i=1}^{K} w_i\,\mathcal{L}_i(\theta), \tag{1}$$

under a KL regularizer of strength β, the abstraction shared by weighted RLHF (Ouyang et al., 2022), fine-grained multi-objective RLHF (Wu et al., 2023), and multi-objective DPO (Zhou et al., 2024).

Optimization strength as a budget. Following Gao et al. (2022), we parameterize optimization strength by the total KL budget D. This budget is set implicitly by the temperature; the correspondence is regime-dependent (Lemma A.9): $D_i \propto \beta_i^{-2}$ for locally linear rewards and $D_i \propto \beta_i^{-1}$ for the log proxy of Eq. (3) in the large-budget regime. The allocation law below is derived for explicit reward maximization with the diminishing-returns (log) proxy, i.e. the β_i^{-1} regime; multi-objective DPO (Algorithm 1) instead instantiates the β_i^{-2} regime, and Section 3 carries the corresponding squared shrinkage and bounded-gain stability separately, so the two regimes are never conflated. For a single objective the proxy rises monotonically in D while gold first rises and then falls—overoptimization (Gao et al., 2022). We ask the multi-objective version: with K objectives sharing one budget, how does the overoptimization split among them?

A minimal model of the phenomenon. We adopt the locally quadratic model of a smooth policy family (Gao et al. (2022); Wang & Huang (2026)): a Gaussian $\pi_\mu = \mathcal{N}(\mu, I)$ around $\pi_0 = \mathcal{N}(0, I)$, so $\mathrm{KL}(\pi_\mu \| \pi_0) = \frac{1}{2}\|\mu\|^2$. Each objective $i \in [K]$ owns a gold direction u_i and a spurious direction v_i (surface features: verbosity, formatting); the u_i are mutually orthonormal, and the v_i are mutually orthonormal and orthogonal to all u_j (the correlated case is stress-tested in Appendix B).

The rewards are

$$g_i(\mu) = a_i \min(\langle u_i, \mu\rangle, C_i) + \tilde{b}_i\,\langle v_i, \mu\rangle - \gamma_i\,\langle v_i, \mu\rangle^2, \tag{2}$$

$$p_i(\mu) = a_i \min(\langle u_i, \mu\rangle, C_i) + b_i \log\big(1 + \langle v_i, \mu\rangle\big). \tag{3}$$

The cap C_i is finite task performance; the hackability b_i is the proxy's spurious sensitivity (zero for exact rules), γ_i the gold fragility, and $\tilde{b}_i$ the benefit of mild spurious movement. Gold is quadratic

in the spurious direction—the documented length–quality curve (Gao et al., 2022; Singhal et al., 2023)—while the proxy is monotone with diminishing returns (Skalse et al., 2022); this asymmetry is the engine of overoptimization.

Assumption 2.1 (Optimization protocol)**.** Training maximizes the weighted proxy

$$\max_{\mu:\ \frac{1}{2}\|\mu\|^2 \le D} \sum_{i=1}^{K} w_i\, p_i(\mu), \qquad w_i > 0, \tag{4}$$

with the budget D growing with optimization strength; the weights are the user's preferences and are fixed during training.

2.2 The Allocation Law

Define the effective hackability

$$c_i \;:=\; \frac{w_i b_i}{\sum_{j=1}^{K} w_j b_j}, \qquad \sum_i c_i = 1. \tag{5}$$

Lemma 2.2 (Spurious budget allocation)**.** Under Assumption 2.1, in the large-budget regime (caps saturated), the optimizer allocates spurious movement

$$v_i(D) \;\approx\; \sqrt{2\big(D - \tfrac{1}{2}\textstyle\sum_j C_j^2\big)\, c_i} \;\approx\; \sqrt{2D\, c_i}, \tag{6}$$

to leading order as $D \to \infty$ (the first "$\approx$" drops the linear term in the first-order condition, valid for $v_i \gg 1$). Equivalently, the spurious KL budget share $D_i = \frac{1}{2}v_i^2$ is proportional to the effective hackability c_i, while the spurious coordinate displacement v_i is proportional to its square root $\sqrt{c_i}$.

Theorem 2.3 (Differential overoptimization / allocation law)**.** Under Assumption 2.1, in the large-budget regime (caps saturated), write $\tilde{D} := D - \frac{1}{2}\sum_j C_j^2$ for the effective spurious budget. To leading order in $\tilde{D}$—i.e. for $\tilde{D} \gg \frac{1}{2}\max_i c_i^{-1}$, where v_i^2 dominates v_i in the first-order condition (Lemma 2.2)—the gold trajectory of objective i is

$$G_i(\tilde{D}) \;=\; a_i C_i + \tilde{b}_i\sqrt{2\tilde{D}c_i} - 2\gamma_i c_i \tilde{D}, \tag{7}$$

and consequently:

(i) (Hump) G_i rises then falls; the post-peak decline $\Delta_i(\tilde{D}) := a_i C_i - G_i(\tilde{D})$ grows as $\Delta_i(\tilde{D}) \approx 2\gamma_i c_i \tilde{D}$;

(ii) (Allocation) total overoptimization splits across objectives as

$$\Delta_i \;\propto\; \gamma_i\, w_i\, b_i, \tag{8}$$

proportional to gold fragility $\times$ user weight $\times$ hackability;

(iii) (Collapse order) the collapse point in effective spurious budget is

$$\tilde{D}_i^\star = \frac{\tilde{b}_i^2}{8\,\gamma_i^2\, c_i}, \qquad \text{total budget } D_i^\star = \tilde{D}_i^\star + \tfrac{1}{2}\textstyle\sum_j C_j^2, \tag{9}$$

smallest for the objective with the largest $\gamma_i^2 c_i/\tilde{b}_i^2$; writing $P_i = \tilde{b}_i^2/(8\gamma_i^2)$, $\tilde{D}_i^\star = P_i/c_i$, so whenever $P_i = P_j$—or, more generally, $P_i/P_j < c_i/c_j$ for every pair with $c_i > c_j$—the ranking is the reverse of the ranking of the c_i's: the most hackable objective (under the stated condition) collapses first.

Corollary 2.4 (Exact-signal immunity, orthogonal channels)**.** For an exact rule-based signal, $b_i = 0$ implies $c_i = 0$: the objective receives no spurious budget ($v_i = 0$), suffers no overoptimization ($\Delta_i = 0$), and its gold quality is monotone non-decreasing in D. Exact signals cannot be over-optimized against gold—within the orthogonal-channel model. On shared channels the no-overoptimization immunity itself is intact (the exact signal's proxy equals its gold, so no proxy–gold gap can open), but the gold-monotonicity part fails when the pinning point $v^\star > 0$: the hackable objective's push on the shared verbosity channel then drives the exact objective's own gold down ($g_2 = a_2 C_2 - \tilde{b}_2 v$ decreases); see Proposition 2.9 and Remark A.6.

Remark 2.5 (What reweighting can and cannot do)**.** Total overoptimization is $\sum_i \Delta_i \approx 2D \cdot (\sum_i \gamma_i w_i b_i)/(\sum_j w_j b_j)$, a weighted average of fragilities, so reweighting changes the total but cannot shrink damage without moving the user's operating point: the shares c_i are the trade-off, and only the budget D contracts every D_i while leaving intent intact (Section 3). Temperature scaling, by its reward-space reading, is a reliability reweighting $\tilde{w}_i = w_i/(1+\eta\hat{g}_i)$ (Proposition A.22)—damage-driven and self-limiting, unlike fixed user reweighting.

Assumption 2.6 (Condition B: held-out spurious response)**.** On the collapse branch ($v_i > \tilde{b}_i/2\gamma_i$), the held-out margin's spurious response $h_i(v_i)$ is strictly decreasing in v_i (equivalently, the held-out margin is a monotone decreasing function of the damage Δ_i), up to noise.

Theorem 2.7 (Bridge decomposition)**.** Extend the model of Section 2 by one memorization direction m_i per objective (unit; the m_i are mutually orthonormal and orthogonal to all u_j and v_j): the training margin additionally earns $\alpha_i \tilde{m}_i$ with $\tilde{m}_i := \langle m_i, \mu\rangle$, and memorization transfers nothing to held-out pairs. Write $\tilde{D} = D - \frac{1}{2}\sum_i C_i^2$ for the effective spurious budget. In the large-budget regime the optimizer spends $\tilde{m}_i \approx w_i\alpha_i\sqrt{2\tilde{D}/\sum_j (w_j\alpha_j)^2}$ on memorization, the spurious coordinate scales as $v_i = \Theta(\tilde{D}^{1/4})$, and the gap between the training (proxy) margin and the held-out margin of objective i,

$$\hat{g}_i \;=\; \text{margin}_i^{\text{train}} - \text{margin}_i^{\text{held-out}}, \tag{10}$$

decomposes as

$$\hat{g}_i \;=\; \underbrace{\alpha_i \tilde{m}_i}_{\text{memorization}} \;+\; \underbrace{b_i \log(1+v_i) - h_i(v_i)}_{\text{spurious miscalibration}}, \tag{11}$$

with h_i the held-out spurious response (derivation in Appendix A). Consequently: (i) the memorization term never saturates; (ii) under Condition 2.6, the miscalibration term is $S_i(v) = b_i \log(1+v) - h_i(v)$, strictly increasing past the gold peak (since $h_i'(v) < 0$ there, Appendix A), hence a monotone tracker of Δ_i. The bridge is therefore derived up to one explicitly identified, falsifiable condition (Condition 2.6)—and (iii) under Condition 2.6 the held-out margin is a monotone decreasing function of the damage on the collapse branch, so the observable stand-in tracks gold damage there (the exact-equality form is sufficient, not necessary). The two regimes are empirically distinguishable on the logit scale (the sigmoid-scale gap saturates regardless; Remark A.4): if $\hat{g}_i$ keeps rising after gold saturates, the memorization term dominates and the gap is a memorization detector; if $\hat{g}_i$ saturates with gold, it tracks gold damage. Protocol P2 decides between the two.

Remark 2.8 (The bridge, made explicit and testable)**.** Condition 2.6 is not a consequence of the base model—that model has no train/held-out split—which is why Theorem 2.7 isolates it as the single hypothesis the bridge rests on; until Protocol P2 passes, the gap is the observable stand-in, not measured gold damage.

2.3 Shared Channels: When an Exact Signal Caps the Hackable Channel

The orthogonal-channel geometry above isolates each objective's spurious channel, but our benchmark instantiates a different one: on TL;DR, the helpfulness proxy rewards verbosity while the conciseness rule penalizes it—one shared channel with opposite signs. The law below shows where Corollary 2.4 stops applying, and that the remedy of Section 3 remains right.

Keep $\pi_\mu = \mathcal{N}(\mu, I)$ and the gold caps, but replace the two orthogonal spurious directions by one shared verbosity coordinate $v = \langle \boldsymbol{v}, \mu\rangle > -1$:

$$p_1 = a_1 \min(\langle u_1, \mu\rangle, C_1) + b_1 \log(1+v), \qquad g_1 = a_1 C_1 + \tilde{b}_1 v - \gamma_1 v^2, \tag{12}$$

$$p_2 = a_2 \min(\langle u_2, \mu\rangle, C_2) - b_2 v, \qquad g_2 = a_2 C_2 - \tilde{b}_2 v, \tag{13}$$

with $b_1, b_2, \tilde{b}_1, \tilde{b}_2, \gamma_1 > 0$ and u_1, u_2 orthogonal to $\boldsymbol{v}$ and each other; objective 2 is the exact length rule, so $b_2 = \tilde{b}_2$ (its proxy equals its gold; unhackable, but not inert). Define the pinning point $v^\star = w_1 b_1/(w_2 b_2) - 1$ and $\tilde{D} = D - \frac{1}{2}\sum_i C_i^2$. This 1D model isolates the spurious-coordinate trade-off under saturated caps (u_1, u_2 pinned at C_1, C_2): it governs the strong-optimization (decline) branch, while the rising limb of helpfulness gold observed at weak optimization ($\beta \gtrsim 20$, Protocol P7) lies outside it and is attributed to pre-saturation learning along the genuine directions u_1, u_2 (Section 4.3).

Proposition 2.9 (Shared-channel pinning)**.** In the large-budget regime (caps saturated), the optimizer's verbosity coordinate is the projection of the pinning point onto the budget interval, $v(D) = \mathrm{clamp}\big(v^\star, -\sqrt{2\tilde{D}}, \sqrt{2\tilde{D}}\big)$ with $v^\star = w_1 b_1/(w_2 b_2) - 1$, and $v(D) \to v^\star$ as $D \to \infty$. Consequently: if $v^\star > 0$ the channel saturates at $v^\star$ and overoptimization is capped; if $v^\star < 0$ the exact rule becomes the engine of the hackable objective's damage (while the exact objective's own gold improves); and helpfulness turns over iff $v^\star$ lies beyond the gold peak, $w_1 b_1/(w_2 b_2) > 1 + \tilde{b}_1/(2\gamma_1)$. The full statements and proof are in Appendix A.

In reward space, GAPOR on this channel discounts objective 1's reward by $s = 1 + \eta\hat{g}_1$, shifting the pinning point to $v^\star(s) = w_1 b_1/(w_2 b_2 s) - 1$; the exact helpfulness-improvement condition, the no-overshoot bound, and the gain range $\eta \le \frac{w_1 b_1}{w_2 b_2 \left(1+\tilde{b}_1/(2\gamma_1)\right)} - 1$ over which GAPOR is guaranteed non-overshooting are in Remark A.7 (Appendix A). That bound holds rigorously for reward-space reliability weighting (RLHF/GRPO); for joint DPO on a shared channel, temperature scaling is not algebraically equivalent to reward discounting (Proposition A.22(iii)), so the bound is a local approximation there.

3 GAPOR: Gap-Adaptive Per-Objective Regularization

This section turns the allocation law and its runtime observable into a remedy. The law identified the per-objective budget as the determinant of damage ($\Delta_i \approx 2\gamma_i D_i$) and the gap of Eq. (10) as its runtime tracker, so the place to act is the budget D_i (the user's weight is intent) and the damage is measurable through $\hat{g}_i$; together these fix one design principle: each objective's temperature should scale with the overoptimization it currently suffers.

Concretely, GAPOR reads the gap $\hat{g}_i$ of Eq. (10) every T steps on a small calibration set and acts on the per-objective temperature β_i. In the separable model of Section 2, the per-objective budget knob in DPO is the temperature (Lemma A.9), so GAPOR replaces the shared β in the per-objective DPO losses (Rafailov et al., 2024) by

$$\boxed{\beta_i \;=\; \beta\left(1 + \eta\,\frac{\hat{g}_i}{s}\right), \qquad \eta \ge 0,\; s > 0\,,} \tag{14}$$

with the gap clipped at zero so that regularization is never relaxed below the baseline. Here η is the loop gain and $s > 0$ is a fixed reference gap scale ($s = 0.3$ throughout) that rescales the effective feedback gain by $1/s$, while T and the calibration size are estimator settings (300 steps and 300 pairs, ablated in Table 20; Appendix D). For weighted RLHF the same principle applies through the reward-space form of Proposition A.22 (Protocol P4).

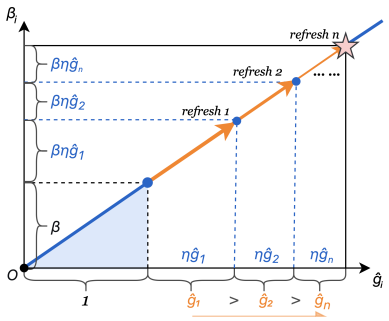

Figure 1: **The geometry of GAPOR.**

Why the temperature is the right knob. The DPO reparameterization $\pi^\star \propto \pi_{\mathrm{ref}} \exp(r_i/\beta_i)$ (Rafailov et al., 2024) makes the per-objective temperature the budget knob in the separable model: in the large-budget (log-proxy) regime $D_i \propto \beta_i^{-1}$, so scaling $\beta_i \leftarrow \beta(1 + \eta\hat{g}_i/s)$ implements

$$D_i \;\approx\; \frac{D_i^\circ}{1 + \eta\hat{g}_i/s}, \qquad D_i^\circ = \tfrac{w_i b_i}{2\beta}, \tag{15}$$

whose regime-corrected correspondence and proof are given in Lemma A.9 (Appendix A). This is the reward-model reading; for the joint DPO loss actually optimized by Algorithm 1 the scaling is $D_i \propto \beta_i^{-2}$ (Remark A.13), and the corresponding squared shrinkage and bounded-gain stability are analyzed separately in Proposition A.18.

Why not the weights. On separable channels scaling β_i equals a reliability reweighting $\tilde{w}_i = w_i/(1 + \eta\hat{g}_i)$ (Proposition A.22(i)–(ii)); on a shared channel the two differ (Proposition A.22(iii)), and the justification is the budget/pinning throttling of Section 2.3. Raw user reweighting moves the preference point permanently; AMoPO (Liu et al., 2025) is closer to the latter and pays a large margin cost on the other objective (Table 1).

What GAPOR provably does. Throughout this subsection the guarantees concern a quasi-static loop: between refreshes the shared policy approaches the optimizer of the current per-objective

Algorithm 1 GAPOR (one-line modification of multi-objective DPO)

Require: weights w, base temperature β, gain η, reference gap scale s, refresh period T, calibration set $\mathcal{D}^{\text{cal}}$

for step $= 1, 2, \ldots$ **do**
  if step $= 1$ **or** step $\bmod T = 0$ **then**
    for each objective i **do**
      $\hat{g}_i \leftarrow \max\big(0,\ \text{margin}_i^{\text{train}} - \text{margin}_i^{\text{cal}}\big)$ {clip at zero}
      $\beta_i \leftarrow \beta\,(1 + \eta\,\hat{g}_i/s)$ {the method}
    end for
  end if
  $\mathcal{L} \leftarrow \sum_i w_i\,\text{DPO-Loss}_i(\beta_i)$
  $\theta \leftarrow \theta - \text{lr}\,\nabla_\theta \mathcal{L}$
end for

budget problem (Assumption 2.1 read with per-objective budgets D_i), and the gap tracks damage through the linear model of Proposition A.14. The quasi-static step is justified by geometric tracking of the reward-model optimizer and by convergence of the (convex) joint DPO loss (Propositions A.10, A.11, A.12). Two regimes must be kept apart: the guarantees above hold in the reward-model β_i^{-1} regime, whereas the DPO loss optimized by Algorithm 1 is in the β_i^{-2} regime, where shrinkage is squared and the loop is stable only below a gain threshold (Remark A.13, Proposition A.18).

Proposition 3.1 (Per-objective defense, reward-model (β_i^{-1}) regime)**.** In the large-budget regime of the reward model of Section 2 (explicit reward maximization, $D_i \propto \beta_i^{-1}$), with memorization subdominant (Remark A.2) and $\hat{g}_i$ treated as fixed between refreshes, GAPOR transforms objective i's overoptimization as

$$\Delta_i^{\text{GAPOR}}(D) \;\approx\; \frac{\Delta_i^{\text{naive}}(D)}{1 + \eta\hat{g}_i/s} \;=\; \frac{2\gamma_i c_i D}{1 + \eta\hat{g}_i/s}, \tag{16}$$

and moves its collapse point from $D_i^\star$ to $D_i^\star\,(1 + \eta\hat{g}_i/s)$. In the DPO ($\beta_i^{-2}$) regime the shrinkage is instead squared, $\Delta_i^{\text{GAPOR}} \approx \Delta_i^{\text{naive}}/(1+\eta\hat{g}_i/s)^2$, and the loop is stable only below a gain threshold (Remark A.13, Proposition A.18).

Proof. Substituting (15) into $\Delta_i \approx 2\gamma_i D_i$ (with $D_i = c_i D$ the per-objective budget; quasi-static: $\hat{g}_i$ fixed between refreshes) gives the shrinkage; the rescaled trajectory $G_i(D) = a_i C_i + \tilde{b}_i\sqrt{2c_i D/(1+\eta\hat{g}_i)} - 2\gamma_i c_i D/(1+\eta\hat{g}_i)$ peaks at $D_i^\star(1+\eta\hat{g}_i)$ (Appendix A). □

Two consequences of this per-objective shrinkage constitute the mechanism. First, the most hacked objective is defended the most: the factor $(1+\eta\hat{g}_i)^{-1}$ uses each objective's own gap, so the collapsing objective receives the largest β_i and the largest collapse-point shift. The allocation ratio is rebalanced: under the loop it becomes

$$\frac{\Delta_i}{\Delta_j} \;\propto\; \frac{\gamma_i w_i b_i}{\gamma_j w_j b_j} \cdot \frac{1+\eta\hat{g}_j}{1+\eta\hat{g}_i}, \tag{17}$$

and, because $\hat{g}_i$ grows with Δ_i (Remark 2.8), the second factor opposes the first. Under the common-slope tracking model $\hat{g}_i = \kappa\Delta_i$ the fixed point is unique and the allocation ratio is strictly reduced below its naive value whenever $\Delta_i > \Delta_j$; the precise statement, its proof, and the caveat that the common slope is an assumption are in Proposition A.14 and Remark A.15 (Appendix A).

Closed-loop convergence. For the reward-model (β_i^{-1}) map, under the same quasi-static model, the refresh map $\Delta^{(t+1)} = A/(1+\eta\kappa\Delta^{(t)})$ with $A = 2\gamma_i c_i D$ converges globally to the balanced fixed point for every gain $\eta\kappa \geq 0$ (contraction for $A\eta\kappa < 1$, monotone via the second iterate otherwise). The full statement, the regime dependence, and the caveat that the exact nonlinear loop need not be globally stable are in Proposition A.16 (the β_i^{-1} map), Proposition A.18 (the β_i^{-2} DPO map), and Remark A.17 (the exact loop; Appendix A).

The signal's practical design (cost, refresh period, and noise stability) is detailed in Appendix D; GAPOR's only extra resource is scoring a 300-pair calibration slice, which adds a few forward passes per refresh and keeps the wall-clock time comparable to the baselines.

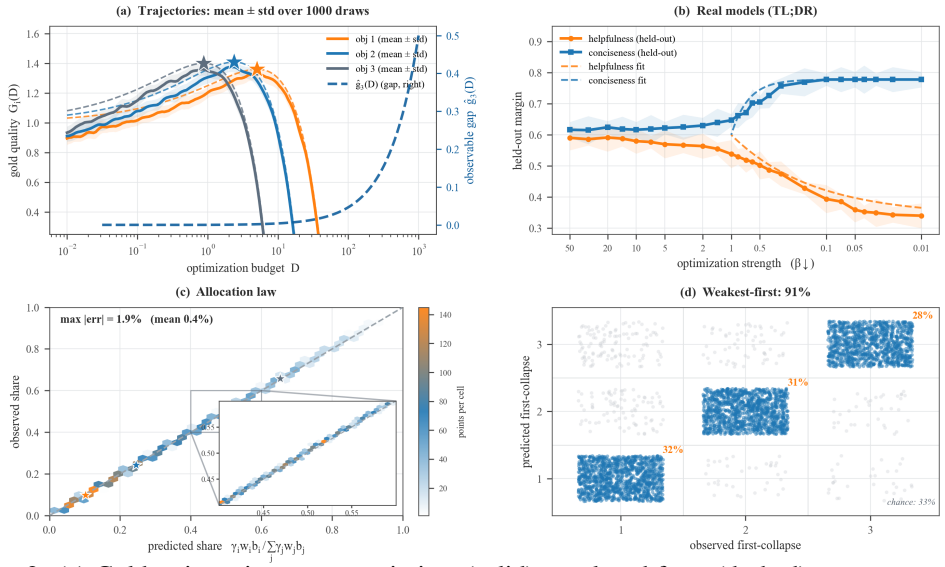

Figure 2: (a) Gold trajectories: exact optimizer (solid) vs. closed form (dashed), stars at predicted collapse points. (b) Real models (TL;DR): held-out helpfulness declines while the exact conciseness signal rises (bands: ±seed std; dashed: monotone-decline guide). (c) Allocation law: observed vs. predicted shares. (d) Weakest-first: predicted vs. observed collapse order.

Table 1: Main comparison at $\beta = 0.03$, equal weights, 10 seeds, held-out proxy margins on 1000 pairs (the within-method damage stand-in of Condition 2.6; cross-method gold is reported in Section 4.5). (paired permutation test, 10^5 resamples, 10 seeds; per-seed values in Table 32). Bold: the only method that closes the gap without a margin cost on the other objective.

Method	**help (held-out)**	**conc (held-out)**	**gap (help)**	**HV**	**vs GAPOR**	
					Δhelp	Δconc
DPO-LW	$0.3923 \pm .0036$	$0.7980 \pm .0170$	$+0.174$	0.5595	$+0.088$	$+0.070$
AMoPO	$0.4220 \pm .0066$	$0.4944 \pm .0089$	$+0.038$	0.4568	$+0.058$	$+0.374$
Constr-RLHF	$0.4419 \pm .0038$	$0.7879 \pm .0158$	$+0.153$	0.5901	$+0.038$	$+0.080$
GAPO	$0.4496 \pm .0045$	$0.8412 \pm .0195$	$+0.174$	0.6149	$+0.031$	$+0.027$
Length-reg. DPO	$0.4377 \pm .0100$	$0.8812 \pm .0150$	$+0.021$	0.6210	$+0.043$	-0.013
GAPOR	$\mathbf{0.4804 \pm .0121}$	$\mathbf{0.8683 \pm .0425}$	$\mathbf{-0.019}$	**0.6456**	–	–

4 Experiments

4.1 Setup

Tasks and objectives. We use the TL;DR summarization comparison benchmark, the standard testbed of DPO-family and multi-objective alignment work (Zhou et al., 2024; Rafailov et al., 2024), as the main task; cross-task evaluation uses the same protocol on HH-RLHF (Bai et al., 2022) and BeaverTails (Ji et al., 2024) (helpfulness/harmlessness; safety/conciseness). On TL;DR, objective 1 is helpfulness (human preference label: noisy, subjective, length-confounded); objective 2 is conciseness (exact length rule).

Models and training. The main grid trains Qwen3-1.7B with multi-objective DPO. The main comparison uses $\beta = 0.03$, a strong-overoptimization regime of Section 2; at the literature-default $\beta = 0.1$ the gap does not open and GAPOR's controller deactivates (Table 18). All methods share data, seeds, and evaluation across Qwen3-1.7B (main), Qwen3-4B, Gemma-3-4B, and Llama-3.2-3B.

Evaluation. We report per-objective held-out preference margins (the damage stand-in of Condition 2.6), the gap $\hat{g}_i$, and the hypervolume proxy $\mathrm{HV} = \sqrt{\mathrm{help} \cdot \mathrm{conc}}$. Every method is run with 10 seeds, and pairwise claims use paired permutation tests over seeds. The gap is a damage statistic, not a quality metric, so margins and HV are reported alongside every gap.

Baselines. **DPO-LW** (the weighted multi-objective DPO template, one shared β); **Constr-RLHF** (Moskovitz et al., 2023) (reward-space CMDP); **AMoPO** (Liu et al., 2025) (reference-free); **GAPO**

Table 2: Cross-model and cross-task generalization ($\beta = 0.03$, equal weights; TL;DR uses a 1000-pair held-out evaluation, HH-RLHF 200 pairs, and BeaverTails the full test split of 33,396 pairs). The gap is the train − held-out margin of the hackable objective (positive = over-driven); bold = smallest $|\text{gap}|$ (closest to zero), the gap-closure criterion applied uniformly across rows. HH-RLHF and BeaverTails use Qwen3-1.7B (TL;DR and BeaverTails use 10 seeds, HH-RLHF 5 seeds; Table 30).

Grid	**objectives**	**gap of obj1 (hackable)**				
		DPO-LW	AMoPO	Constr-RLHF	GAPO	**GAPOR**
Qwen3-4B (TL;DR)	help / conc	+0.212	+0.214	+0.163	+0.075	**−0.011**
Llama-3.2-3B (TL;DR)	help / conc	+0.024	+0.031	+0.016	**+0.007**	−0.056
Gemma-3-4B (TL;DR)	help / conc	**+0.093**	+0.101	+0.166	+0.156	−0.120
HH-RLHF (Qwen3-1.7B)	help / harm	+0.096	+0.067	+0.045	−0.091	**−0.021**
BeaverTails (Qwen3-1.7B)	safety / conc	+0.297	+0.256	+0.177	+0.088	**+0.033**

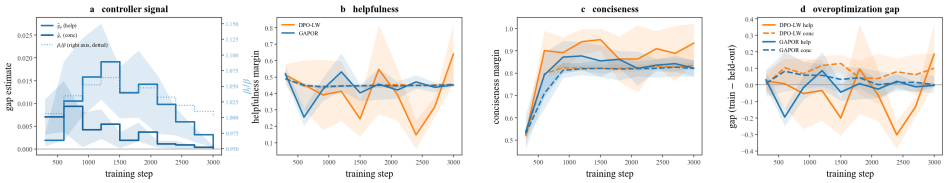

Figure 3: Training dynamics (TL;DR, $\beta = 0.03$, 10 seeds, mean±std): (a) detect–act–settle; (b,c) train vs. held-out margins; (d) per-objective gaps.

(Li et al., 2025) (gradient rebalancing); and **Length-regularized DPO** (Park et al., 2024) (the published length-channel defense). Two further families are discussed in Appendix C but not re-run: **CAGrad** (Liu et al., 2021) (multi-gradient conflict resolution, covered by GAPO) and **Rewarded Soups/WARP** (Ramé et al., 2023; 2024) (post-hoc weight interpolation that never touches the training budget). All baselines run their published algorithms with original objectives; the DPO-based methods (DPO-LW, GAPOR, Length-regularized DPO) share a prompt-free, length-normalized scoring scaffold (Remark D.1), while AMoPO, GAPO, and Constr-RLHF use their original objectives.

4.2 Verification

Synthetic verification. We instantiate the model with three ranked-hackability objectives, sweep the budget, and measure post-peak declines. The law holds (Figure 2): trajectories trace the predicted humps, allocation shares match the predicted diagonal (max error 1.9%), and weakest-first order matches in 91% of draws (robust to misspecification, Appendix B).

Real-model verification. On Qwen3-1.7B/TL;DR the two objectives follow their predicted opposite fates (Figure 2b): held-out helpfulness plateaus then declines, while conciseness rises and saturates (Table 15). With a deliberately under-trained reward model as objective 1, GAPOR suppresses the proxy climb while gold stalls (Table 29).

4.3 The Shared-Channel Falsification Check

Proposition 2.9 governs the saturated regime—the decline branch and its anti-correlation with the exact signal—while the pre-saturation rising limb is genuine task learning along u_1, u_2, outside the 1D model. Table 15 shows the plateau and decline (held-out helpfulness flat at $\beta \geq 1$, declining below $\beta \approx 0.3$) while conciseness rises (0.721 → 0.810). The extended sweep supplies the rising limb: gold helpfulness rises from 3.18 ($\beta = \infty$) to 3.95 at $\beta_{\text{peak}} \approx 20$, then declines to 2.63 ($\beta = 0.01$, Table 13). Attributing both limbs to one shared coordinate v would be inconsistent (Proposition 2.9(ii)–(iii)). Fitting Eq. (22) to the 15 gold points (Tables 13,22) reproduces the hump ($R^2 \approx 0.93$) and confirms the criterion ($v^\star/v_{\text{peak}} \approx 2.3 > 1$); the individual parameters carry a one-parameter scale ambiguity (Remark A.8).

4.4 Main Comparison

GAPOR realizes exactly the combination the law predicts (Table 1): it removes the hackable objective's gap while raising its held-out helpfulness, without degrading the conciseness margin, and attains the highest HV; length-controlled evaluation shows this is not a length artifact (LC margin

Table 3: Same gap signal, different actuators ($\beta = 0.03$, 5 seeds, the protocol of the ablation in Table 20); p from a paired permutation test over the 5 seeds on HV (the 2^5=32 sign assignments give a minimum achievable one-sided p of $1/32$=0.031; all rows use the same one-sided test). The GAPOR row is the 5-seed ablation default (its 10-seed value in Table 1 is HV $= 0.6456$). The static oracle uses the dev-selected ratio $\beta_{\text{help}} = 2\beta$; full isolation baselines in Appendix F. All actuators share the same gap signal and hence the same 11×500-pair calibration scoring; only the tuned-global-β and static-oracle rows additionally query the dev split.

Method	help (held-out)	conc (held-out)	gap (help)	HV	p (vs. ours)
Early stopping	0.4601	0.8563	+0.006	0.6277	0.031
Tuned global β grid	0.4490	0.8060	+0.031	0.6016	0.031
Static per-objective β	0.4612	0.8620	−0.005	0.6305	0.031
Weight clamping	0.4788	0.8862	−0.018	0.6514	0.031
GAPOR	**0.4839**	**0.8911**	**−0.022**	**0.6567**	—

Table 4: Weighted RLHF with GRPO (TL;DR, Qwen3-1.7B, KL $\beta = 0.03$, 5 seeds). GAPOR (reward-scaled) multiplies the helpfulness reward by $(1 + \eta\hat{g}_i)^{-1}$ (Proposition A.22) with the same 200-train/300-calibration gap scoring; gold = 100-prompt judge (1–5). Full baselines in Table 27.

Method	help (held-out)	conc (held-out)	gap (help)	HV	gold help
Weighted RLHF	$0.412 \pm .011$	$0.798 \pm .018$	+0.152	0.5734	$2.9 \pm .10$
GAPOR	$\mathbf{0.468 \pm .010}$	$\mathbf{0.852 \pm .020}$	**−0.011**	**0.6315**	$\mathbf{3.3 \pm .09}$

+0.067, LC-WR 58.2%, Table 9). The pattern transfers across models and tasks (Table 2); on HH-RLHF GAPO's smaller |gap| comes at the cost of the lowest helpfulness margin (Table 30), yet GAPOR's gold helpfulness leads (3.38 vs. DPO-LW's 3.10, Table 12). Training shows the intended detect–act–settle profile (Figure 3). The effect is regime-consistent—present at $\beta = 0.05$, vanishing at $\beta = 0.1$ where the gap no longer opens (Table 18)—so GAPOR is self-limiting and doubles as an automatic regime detector. Ablations attribute the effect to the feedback signal rather than the actuator: every non-temperature actuator achieves lower HV than GAPOR (Table 3), and the gain sweep over-regularizes monotonically without oscillation, consistent with Proposition A.16 (reward-model β_i^{-1}) and Proposition A.18 (DPO β_i^{-2}) below its gain threshold. The one-line principle transfers to reward-space RLHF with GRPO (Table 4) and replicates at 7B with 10-seed gold (Table 10: 3.31 vs. 2.95, $p < 0.001$) and at every user weight (Figure 5).

4.5 Independent Gold Evaluation

We score every sweep checkpoint and the five $\beta = 0.03$ policies on 250 held-out prompts with a two-axis 1–5 rubric (judge–human $\rho = 0.72$, Table 24). GAPOR's gold helpfulness declines more slowly than the baseline's ($3.34 \pm .06$ vs. $2.91 \pm .06$ at $\beta = 0.03$, paired $p < 0.001$), with gold conciseness unchanged ($p = 0.3$), so GAPOR protects gold, not merely the proxy; it beats every baseline on helpfulness (win rates 57.8–62.4%) and ties on conciseness (Table 23). The logit-scale discriminator confirms miscalibration dominance ($\hat{g}_i/\Delta_i \approx 1.14$, Table 14). Length control leaves GAPOR's advantage intact (Table 9), and the gold leads hold across grids (7B 3.31 vs. 2.95, $p < 0.001$; K=3 3.42 vs. 3.05; HH-RLHF 3.38 vs. 3.10; Tables 10,11,12).

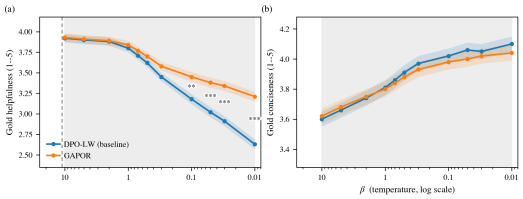

Figure 4: Absolute gold curves along the β-sweep (250 held-out prompts): (a) gold helpfulness—the baseline declines monotonically while GAPOR attenuates the decline; (b) gold conciseness—no measurable trade-off.

5 Conclusion

We derived a closed-form allocation law for multi-objective overoptimization, identified the train–held-out gap as its runtime observable, and built GAPOR, with provable shrinkage, rebalancing, and closed-loop convergence that is global in the reward-model regime and bounded-gain in DPO, as well as gold protection verified by independent evaluation. The theory is falsified if gold fails to exhibit a hump, if the allocation ratio does not track $\gamma_i w_i b_i$, or if the collapse criterion fails on gold.

AI USE STATEMENT

Language models were only used for language polishing. All scientific content was produced and verified by the authors.

REFERENCES

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. A general theoretical paradigm to understand learning from human preferences, 2023. URL https://arxiv.org/abs/2310.12036.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

Lichang Chen, Chen Zhu, Davit Soselia, Jiuhai Chen, Tianyi Zhou, Tom Goldstein, Heng Huang, Mohammad Shoeybi, and Bryan Catanzaro. Odin: Disentangled reward mitigates hacking in rlhf, 2024. URL https://arxiv.org/abs/2402.07319.

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 (ICML)*, 2018.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2017. URL https://arxiv.org/abs/1706.03741.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback, 2023. URL https://arxiv.org/abs/2310.12773.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators, 2024. URL https://arxiv.org/abs/2404.04475.

Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alex D'Amour, DJ Dvijotham, Adam Fisch, Katherine Heller, Stephen Pfohl, Deepak Ramachandran, Peter Shaw, and Jonathan Berant. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking, 2024. URL https://arxiv.org/abs/2312.09244.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization, 2024. URL https://arxiv.org/abs/2402.01306.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization, 2022. URL https://arxiv.org/abs/2210.10760.

Yiju Guo, Ganqu Cui, Lifan Yuan, Ning Ding, Zexu Sun, Bowen Sun, Huimin Chen, Ruobing Xie, Jie Zhou, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Controllable preference optimization: Toward controllable multi-objective alignment, 2024. URL https://arxiv.org/abs/2402.19085.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

Jacek Karwowski, Oliver Hayman, Xingjian Bai, Klaus Kiendlhofer, Charlie Griffin, and Joar Skalse. Goodhart's law in reinforcement learning, 2023. URL https://arxiv.org/abs/2310.09144.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Dahyun Kim, Yungi Kim, Wonho Song, Hyeonwoo Kim, Yunsu Kim, Sanghoon Kim, and Chanjun Park. sdpo: Don't use your data all at once, 2024. URL https://arxiv.org/abs/2403.19270.

Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. Specification gaming: the flip side of ai ingenuity. DeepMind Blog, 2020.

Chengao Li, Hanyu Zhang, Yunkun Xu, Hongyan Xue, Xiang Ao, and Qing He. Gradient-adaptive policy optimization: Towards multi-objective alignment of large language models, 2025. URL https://arxiv.org/abs/2507.01915.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning, 2021. URL https://arxiv.org/abs/2110.14048.

Qi Liu, Jingqing Ruan, Hao Li, Haodong Zhao, Desheng Wang, Jiansong Chen, Wan Guanglu, Xunliang Cai, Zhi Zheng, and Tong Xu. Amopo: Adaptive multi-objective preference optimization without reward models and reference models, 2025. URL https://arxiv.org/abs/2506.07165.

Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J. Liu, and Jialu Liu. Statistical rejection sampling improves preference optimization, 2023. URL https://arxiv.org/abs/2309.06657.

David Manheim and Scott Garrabrant. Categorizing variants of goodhart's law, 2018. URL https://arxiv.org/abs/1803.04585.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward, 2024. URL https://arxiv.org/abs/2405.14734.

Ted Moskovitz, Aaditya K. Singh, DJ Strouse, Tuomas Sandholm, Ruslan Salakhutdinov, Anca D. Dragan, and Stephen McAleer. Confronting reward model overoptimization with constrained rlhf, 2023. URL https://arxiv.org/abs/2310.04373.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. 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 Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization, 2024. URL https://arxiv.org/abs/2403.19159.

Lutz Prechelt. Early stopping - but when? In *Neural Networks: Tricks of the Trade*. Springer, 1998.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

Alexandre Ramé, Guillaume Couairon, Mustafa Shukor, Corentin Dancette, Jean-Baptiste Gaya, Laure Soulier, and Matthieu Cord. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards, 2023. URL https://arxiv.org/abs/2306.04488.

Alexandre Ramé, Johan Ferret, Nino Vieillard, Robert Dadashi, Léonard Hussenot, Pierre-Louis Cedoz, Pier Giuseppe Sessa, Sertan Girgin, Arthur Douillard, and Olivier Bachem. Warp: On the benefits of weight averaged rewarded policies, 2024. URL https://arxiv.org/abs/2406.16768.

D. M. Roijers, P. Vamplew, S. Whiteson, and R. Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113, October 2013. ISSN 1076-9757. doi: 10.1613/jair.3987. URL http://dx.doi.org/10.1613/jair.3987.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Disha Singha. Uncertainty-aware reward discounting for mitigating reward hacking, 2026. URL https://arxiv.org/abs/2604.26360.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in rlhf, 2023. URL https://arxiv.org/abs/2310.03716.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming, 2022. URL https://arxiv.org/abs/2209.13085.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback, 2022. URL https://arxiv.org/abs/2009.01325.

Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. Beyond reverse kl: Generalizing direct preference optimization with diverse divergence constraints, 2023. URL https://arxiv.org/abs/2309.16240.

Jiacheng Wang and Jinbin Huang. Reward hacking as equilibrium under finite evaluation, 2026. URL https://arxiv.org/abs/2603.28063.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training, 2023. URL https://arxiv.org/abs/2306.01693.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, pp. 100211, 2024.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning, 2020. URL https://arxiv.org/abs/2001.06782.

Yifan Zhong, Chengdong Ma, Xiaoyuan Zhang, Ziran Yang, Haojun Chen, Qingfu Zhang, Siyuan Qi, and Yaodong Yang. Panacea: Pareto alignment via preference adaptation for llms, 2024. URL https://arxiv.org/abs/2402.02030.

Zhanhui Zhou, Jie Liu, Jing Shao, Xiangyu Yue, Chao Yang, Wanli Ouyang, and Yu Qiao. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization, 2024. URL https://arxiv.org/abs/2310.03708.

Simon Zhuang and Dylan Hadfield-Menell. Consequences of misaligned ai, 2020. URL https://arxiv.org/abs/2102.03896.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020. URL https://arxiv.org/abs/1909.08593.

A Missing Proofs and Omitted Statements

The extremal-regime statement and the proofs deferred from Section 2 for space are collected here. We first restate the setup in the notation of Section 2, so that this appendix is self-contained. The policy is a Gaussian $\pi_\mu = \mathcal{N}(\mu, I)$ around the reference $\pi_0 = \mathcal{N}(0, I)$, so $\mathrm{KL}(\pi_\mu \| \pi_0) = \frac{1}{2}\|\mu\|^2$ and the budget constraint reads $\frac{1}{2}\|\mu\|^2 \le D$. Objective i owns a gold direction u_i and a spurious direction v_i; the u_i are mutually orthonormal, and the v_i are mutually orthonormal and orthogonal to all u_j, with rewards as in (2)–(3). The effective hackability is $c_i = w_i b_i / \sum_j w_j b_j$, and the optimizer solves $\max_{\mu:\, \frac{1}{2}\|\mu\|^2 \le D} \sum_i w_i p_i(\mu)$. The large-budget regime means every gold cap is saturated, i.e., the optimizer sets $\langle u_i, \mu \rangle = C_i$ for all i (there is no marginal benefit beyond the cap, and the regime assumes D is large enough that the caps are reached).

Proposition A.1 (Extremal regime)**.** In the limit $w_i \to 1$, $c_i \to 1$, so the collapse point of objective i tends to $\tilde{b}_i^2/(8\gamma_i^2)$ and the decline slope Δ_i/D tends to $2\gamma_i$: the heavily weighted objective becomes the first to collapse under its own optimization, the quantitative form of extremal Goodhart (Manheim & Garrabrant, 2018).

Proof. Fix objective i and take $w_i \to 1$ with the remaining weights $w_j \to 0$ for $j \neq i$ (normalized, $\sum_k w_k = 1$). Substituting into (5),

$$c_i \;=\; \frac{w_i b_i}{w_i b_i + \sum_{j \neq i} w_j b_j} \;\longrightarrow\; 1, \qquad c_j \;\longrightarrow\; 0 \;\; (j \neq i).$$

By (9), $D_i^\star \to \tilde{b}_i^2/(8\gamma_i^2)$, which is finite, while $D_j^\star \to \infty$ for every $j \neq i$: objective i is therefore the unique first collapse. By (8) its decline slope is $\Delta_i/D \to 2\gamma_i c_i \to 2\gamma_i$, the maximal slope available to that objective. Both limits are the quantitative form of extremal Goodhart: the more the user weights an objective, the earlier and the faster its own optimization destroys it. □

Proof of Lemma 2.2. Write $v_i = \langle v_i, \mu \rangle$ for the spurious coordinate and form the Lagrangian of (4),

$$\Phi(\mu, \lambda) \;=\; \sum_i w_i p_i(\mu) \;-\; \lambda\Big(\tfrac{1}{2}\|\mu\|^2 - D\Big), \qquad \lambda \ge 0.$$

Because the v_i are mutually orthogonal and orthogonal to all u_j, each spurious coordinate decouples, and the first-order condition along v_i is

$$\frac{\partial \Phi}{\partial v_i} \;=\; \frac{w_i b_i}{1 + v_i} - \lambda v_i \;=\; 0 \qquad \Longleftrightarrow \qquad v_i^2 + v_i \;=\; \frac{w_i b_i}{\lambda}.$$

The multiplier is positive: every spurious direction has positive marginal proxy gain ($w_i b_i/(1+v_i) > 0$ whenever $b_i > 0$) and positive marginal KL cost (λv_i), so the optimum lies on the boundary $\frac{1}{2}\|\mu\|^2 = D$ and $\lambda > 0$. In the large-budget regime the gold coordinates sit at their caps, contributing the constant $\sum_i w_i a_i C_i$ to the objective; the budget constraint therefore reads

$$\tfrac{1}{2}\sum_i v_i^2 \;=\; D - \tfrac{1}{2}\sum_i C_i^2 \;=\; D - \text{const}.$$

It remains to solve (A). For large D, each v_i is large and v_i^2 dominates v_i, so $v_i^2 \approx w_i b_i/\lambda$, i.e., $v_i \approx \sqrt{w_i b_i/\lambda}$. Substituting into the budget,

$$\tfrac{1}{2}\sum_i \frac{w_i b_i}{\lambda} \;\approx\; D - \tfrac{1}{2}\sum_j C_j^2 \qquad \Longrightarrow \qquad \lambda^{-1} \sum_i w_i b_i \;\approx\; 2\Big(D - \tfrac{1}{2}\sum_j C_j^2\Big),$$

hence $\lambda^{-1} \approx 2\big(D - \frac{1}{2}\sum_j C_j^2\big)/\sum_j w_j b_j$ and

$$v_i \;\approx\; \sqrt{2\Big(D - \tfrac{1}{2}\sum_j C_j^2\Big)\,\frac{w_i b_i}{\sum_j w_j b_j}} \;=\; \sqrt{2\Big(D - \tfrac{1}{2}\sum_j C_j^2\Big)\,c_i}\,,$$

which is (6) (and $\approx \sqrt{2Dc_i}$ in the large-budget regime). Finally, the boundary case $b_i = 0$ gives $v_i = 0$ in (A): a direction with no proxy sensitivity earns nothing and spends nothing. □

Proof of Theorem 2.3. Part (i). Substitute (6) into (2) with the cap saturated, writing $\tilde{D} = D - \frac{1}{2}\sum_j C_j^2$:

$$G_i(\tilde{D}) \;=\; a_i C_i + \tilde{b}_i\sqrt{2\tilde{D}c_i} - \gamma_i(2\tilde{D}c_i) \;=\; a_i C_i + \tilde{b}_i\sqrt{2c_i}\,\sqrt{\tilde{D}} - 2\gamma_i c_i \tilde{D}.$$

Differentiating in $\tilde{D}$,

$$\frac{\partial G_i}{\partial \tilde{D}} \;=\; \tilde{b}_i\sqrt{\frac{c_i}{2}}\,\tilde{D}^{-1/2} - 2\gamma_i c_i, \qquad \frac{\partial^2 G_i}{\partial \tilde{D}^2} \;=\; -\frac{\tilde{b}_i}{2}\sqrt{\frac{c_i}{2}}\,\tilde{D}^{-3/2} < 0,$$

so G_i is strictly concave in $\tilde{D}$: it rises (the derivative diverges to $+\infty$ as $\tilde{D} \to 0^+$) and then falls (the derivative tends to $-2\gamma_i c_i < 0$). The unique peak solves $\tilde{b}_i\sqrt{c_i/2}\,\tilde{D}^{-1/2} = 2\gamma_i c_i$, i.e., $\tilde{D}_i^\star = \tilde{b}_i^2/(8\gamma_i^2 c_i)$, which is (9); in total budget $D_i^\star = \tilde{D}_i^\star + \frac{1}{2}\sum_j C_j^2$. Beyond the peak, $\Delta_i(\tilde{D}) := a_i C_i - G_i(\tilde{D}) = 2\gamma_i c_i \tilde{D} - \tilde{b}_i\sqrt{2c_i}\sqrt{\tilde{D}}$, whose linear term dominates for $\tilde{D} \gg \tilde{D}_i^\star$, giving the decline $\Delta_i(\tilde{D}) \approx 2\gamma_i c_i \tilde{D}$.

Part (ii). From the decline, $\Delta_i \approx 2\gamma_i c_i \tilde{D}$; the denominator of (5) is common to all objectives, so $\Delta_i \propto \gamma_i w_i b_i$, which is (8).

Part (iii). Write $P_i = \tilde{b}_i^2/(8\gamma_i^2)$, so $\tilde{D}_i^\star = P_i/c_i$ is strictly decreasing in c_i at fixed P_i. The ranking of $\tilde{D}_i^\star$ is the reverse of the ranking of the c_i's whenever $P_i = P_j$ for all pairs—or, more generally, whenever $P_i/P_j < c_i/c_j$ holds for every pair with $c_i > c_j$, since then $P_i/c_i < P_j/c_j$. Under equal weights $c_i \propto b_i$, so the objective with the largest b_i—the most hackable one, under the stated condition on the P_i's—collapses first. □

Proof of Corollary 2.4. If $b_i = 0$ (and at least one other objective has $b_j > 0$, so that $\sum_j w_j b_j > 0$), then $c_i = 0$ by (5), and (6) gives $v_i(D) = 0$ at every budget D: the exact signal receives no spurious movement. Its gold trajectory is then constant after saturation, $G_i(D) = a_i C_i$ for all D, hence monotone non-decreasing, and $\Delta_i(D) = a_i C_i - G_i(D) = 0$ at every D. (If every objective had $b_i = 0$, no spurious direction would exist at all and the statement is vacuous.) This argument uses orthogonality; on shared channels the no-overoptimization immunity is intact (proxy = gold, so no proxy–gold gap opens), but the gold-monotonicity conclusion fails in the quantified way of Proposition 2.9 (Remark A.6). □

Full proof of Theorem 2.7 (bridge decomposition). The training margin adds a memorization channel with reward $\alpha_i \tilde{m}_i$; the objective weights w_i enter the first-order conditions along v_i and m_i,

$$\frac{w_i b_i}{1 + v_i} = \lambda v_i, \qquad w_i \alpha_i = \lambda \tilde{m}_i, \tag{18}$$

and the budget reads $\frac{1}{2}(\sum_i v_i^2 + \sum_i \tilde{m}_i^2) = \tilde{D}$. For large $\tilde{D}$, $\frac{1}{2}(\sum_i w_i b_i/\lambda + \sum_i (w_i\alpha_i)^2/\lambda^2) = \tilde{D}$, so the memorization term dominates the budget and $\lambda \approx \sqrt{\sum_j (w_j\alpha_j)^2/(2\tilde{D})}$, $\tilde{m}_i = w_i\alpha_i/\lambda \approx w_i\alpha_i\sqrt{2\tilde{D}/\sum_j (w_j\alpha_j)^2}$, and $v_i^2 \approx w_i b_i/\lambda = w_i b_i\sqrt{2\tilde{D}/\sum_j (w_j\alpha_j)^2}$, i.e. $v_i = \Theta(\tilde{D}^{1/4})$. The train margin is $b_i \log(1+v_i) + \alpha_i \tilde{m}_i$ (plus the saturated cap term) and the held-out margin is $h_i(v_i)$, which is Eq. (11). Under Condition 2.6, $h_i'(v) < 0$ for $v > \tilde{b}_i/(2\gamma_i)$, so $S_i(v) = b_i \log(1+v) - h_i(v)$ with $S_i'(v) = b_i/(1+v) - h_i'(v) > b_i/(1+v) > 0$ there; likewise $\Delta_i'(v) = 2\gamma_i v - \tilde{b}_i > 0$. Both S_i and Δ_i are strictly increasing in v—and $v_i(\tilde{D})$ is increasing—so S_i is a monotone tracker of Δ_i on the collapse branch, and the memorization term $\alpha_i \tilde{m}_i \approx w_i \alpha_i^2\sqrt{2\tilde{D}/\sum_j (w_j\alpha_j)^2}$ never saturates, giving the two discriminable regimes of the theorem. □

Remark A.2 (Regime bookkeeping: the memorization extension changes the damage scaling)**.** The allocation law of Theorem 2.3 is derived without memorization, where $v_i = \Theta(D^{1/2})$ and $\Delta_i \approx 2\gamma_i c_i D$ grows linearly in D. Adding the memorization directions changes this: the memorization term dominates the budget (λ^{-2} vs. λ^{-1} in Eq. (18)), so the spurious coordinate is compressed to $v_i = \Theta(\tilde{D}^{1/4})$ and the spurious damage obeys $\Delta_i = \gamma_i v_i^2 - \tilde{b}_i v_i = \Theta(\tilde{D}^{1/2})$. The two models therefore describe two regimes: the allocation law (and the defense of Section 3, which builds on

$\Delta_i \approx 2\gamma_i c_i D$) is the operative one when memorization is subdominant, i.e. when the miscalibration term of Eq. (11) dominates the gap; the memorization-dominated regime is the one where the gap must be re-read as a memorization detector rather than a damage tracker (Theorem 2.7). Protocol P2 decides between the two empirically.

Remark A.3 (Memorization vs. miscalibration shares of the gap)**.** In the large-$\tilde{D}$ limit with the quadratic held-out response $h_i(v) = \tilde{b}_i v - \gamma_i v^2$ (the sharp form of Condition 2.6), the noise-free diagonal-budget model of the proof gives closed-form shares for each objective. Writing $\hat{m}_i^{\mathrm{mem}} = \alpha_i \tilde{m}_i$ and $\hat{m}_i^{\mathrm{cal}} = b_i \log(1+v_i) - h_i(v_i)$, one has, for fixed coefficients with $\gamma_i b_i > 0$ and some $\alpha_j \neq 0$,

$$\frac{\hat{g}_i}{\Delta_i} \;\to\; 1 + \frac{\alpha_i^2}{\gamma_i b_i}, \qquad \frac{\hat{m}_i^{\mathrm{mem}}}{\hat{g}_i} \;\to\; \frac{\alpha_i^2}{\alpha_i^2 + \gamma_i b_i} \qquad (\tilde{D} \to \infty), \tag{19}$$

so a gap that follows damage can nonetheless be dominated by memorization (e.g. $\alpha_i^2 = 9\gamma_i b_i$ puts 90% of the gap in the memorization term while the gap tracks ten times the damage). Because the slopes $\alpha_i^2/(\gamma_i b_i)$ are objective-specific, Eq. (19) also shows that the gap–damage slope is not common across objectives in general; a common tracking slope κ (Proposition A.14) is an additional modeling assumption, not a consequence of Theorem 2.7. The protocol P2 therefore reports both the correlation and, where estimable, the memorization share, rather than concluding dominance from the correlation alone.

Remark A.4 (The discriminator, and why the sigmoid score is the wrong scale)**.** Eq. (19) supplies an explicit, identifiable discriminator: the gap tracks damage (miscalibration-dominated) iff $\hat{g}_i/\Delta_i \approx 1$, i.e. $\alpha_i^2 \ll \gamma_i b_i$, whereas a memorization-dominated gap has $\hat{g}_i/\Delta_i \approx 1 + \alpha_i^2/(\gamma_i b_i) \gg 1$. The ratio identifies the combination $\alpha_i^2/(\gamma_i b_i)$—hence the memorization share $\alpha_i^2/(\alpha_i^2+\gamma_i b_i)$—without separately identifying α_i, b_i, γ_i. Two caveats. (i) Scale. The ratio must be computed on the logit scale: the reported margins are mean sigmoid scores $\sigma(z)$, which saturate as the underlying logit z grows, so the sigmoid-scale gap obeys $0 \le \hat{g}_i/\Delta_i \le 1/\Delta_i \to 0$ even when memorization dominates; reporting the mean log-ratio z (already the argument of the sigmoid inside the DPO loss) instead of the mean $\sigma(z)$ removes this compression. (ii) Data. The ratio requires the damage Δ_i, hence an independent gold signal; it is therefore computed on the task where gold is available (TL;DR, Protocol P2). The saturation-based test of Theorem 2.7 is a coarse proxy of this ratio, valid only on the logit scale; the computed ratio is reported in Table 14 (Protocol P8).

Remark A.5 (Closed form for a family of diminishing-returns proxies)**.** Take $p_i = b_i v_i^{\alpha}$ with $\alpha \in (0,1)$. The first-order condition along v_i is $w_i b_i \alpha v_i^{\alpha-1} = \lambda v_i$, i.e. $v_i = (w_i b_i \alpha/\lambda)^{1/(2-\alpha)}$; the budget $\frac{1}{2}\sum_i v_i^2 = \tilde{D}$ gives $\lambda = \big(S_\alpha/(2\tilde{D})\big)^{(2-\alpha)/2}$ with $S_\alpha = \sum_j (w_j b_j \alpha)^{2/(2-\alpha)}$, so $v_i = (w_i b_i \alpha)^{1/(2-\alpha)} (2\tilde{D}/S_\alpha)^{1/2}$ and $\Delta_i \approx \gamma_i v_i^2 \propto \gamma_i (w_i b_i)^{2/(2-\alpha)}$. As $\alpha \to 0$ the exponent tends to 1, recovering Eq. (8) (the log proxy of (3) is the $\alpha \to 0$ limit of the rescaled family $b_i(v^\alpha - 1)/\alpha \to b_i \log v$); as $\alpha \to 1$ (linear proxy) the exponent tends to 2, the super-concentration observed in the misspecification study (Table 5). The collapse point is likewise closed-form: $v_i(D) = \tilde{b}_i/(2\gamma_i)$ at $D_i^\star = \frac{1}{2} S_\alpha \big(\tilde{b}_i/(2\gamma_i (w_i b_i \alpha)^{1/(2-\alpha)})\big)^2$. The qualitative statement—damage concentrates on the most hackable objective—is therefore structural for the whole diminishing-returns family, and only the quantitative exponent depends on α.

Remark A.6 (Full statement of Proposition 2.9)**.** With the setup of Section 2.3, the three consequences of the clamp are:

(i) **(Capping)** If $v^\star > 0$, the channel saturates at $v^\star$ once $\tilde{D} \ge v^{\star 2}/2$, capping overoptimization at $\Delta_1(v^\star) = \gamma_1 v^{\star 2} - \tilde{b}_1 v^\star$ for all budgets; if $v^\star \le \tilde{b}_1/(2\gamma_1)$, helpfulness gold is monotone in D.

(ii) **(The exact signal becomes the engine of the hackable objective's damage)** If $v^\star < 0$, the optimizer drives v toward $v^\star < 0$ and helpfulness degrades—$g_1 \to a_1 C_1 + \tilde{b}_1 v^\star - \gamma_1 v^{\star 2} < a_1 C_1$—while the exact objective's own gold improves: $g_2(v^\star) - g_2(0) = -\tilde{b}_2 v^\star > 0$. The exact signal's immunity is intact; what fails is the hackable objective's protection (Corollary 2.4 is orthogonal-channel only).

(iii) **(Falsifiable collapse criterion)** Helpfulness gold turns over iff $v^\star$ lies beyond the gold peak, $w_1 b_1/(w_2 b_2) > 1 + \tilde{b}_1/(2\gamma_1)$; net damage requires $v^\star > \tilde{b}_1/\gamma_1$.

Full proof of Proposition 2.9. With caps saturated, the problem in v alone is $\max f(v) = w_1 b_1 \log(1+v) - w_2 b_2 v$ subject to $|v| \le \sqrt{2\tilde{D}}, v > -1$. f is strictly concave with unique maximizer $v^\star = w_1 b_1/(w_2 b_2) - 1$ (note $v^\star > -1$ always, so the log domain is respected), and the constrained maximizer is the projection of $v^\star$ onto the interval—the clamp. (i): $\min(v^\star, \sqrt{2\tilde{D}}) = v^\star$ once $\tilde{D} \ge v^{\star 2}/2$; substitute into g_1; monotonicity follows since g_1 increases on $[0, \tilde{b}_1/(2\gamma_1)]$ and $v(D)$ increases from 0 to $v^\star$ along the budget. (ii): $v \to v^\star < 0$, so $\tilde{b}_1 v^\star - \gamma_1 v^{\star 2} < 0$ (objective 1 degrades) while $g_2(v^\star) - g_2(0) = -\tilde{b}_2 v^\star > 0$ (objective 2 improves). (iii): solve $v^\star > \tilde{b}_1/(2\gamma_1)$ for the turning point, and $v^\star > \tilde{b}_1/\gamma_1$ for net damage. □

Remark A.7 (GAPOR on the shared channel, in reward space)**.** Discounting objective 1's reward by $s_{\text{disc}} = 1 + \eta \hat{g}_1$ replaces the objective $f(v) = w_1 b_1 \log(1+v) - w_2 b_2 v$ by $f_{s_{\text{disc}}}(v) = w_1 b_1 \log(1+v)/s_{\text{disc}} - w_2 b_2 v$, whose unique maximizer is the shifted pinning point $v^\star(s_{\text{disc}}) = w_1 b_1/(w_2 b_2 s_{\text{disc}}) - 1$, a strictly decreasing function of s_{disc}. Writing $g_1(v) = a_1 C_1 + \gamma_1 v_{\text{pk}}^2 - \gamma_1 (v - v_{\text{pk}})^2$ with $v_{\text{pk}} = \tilde{b}_1/(2\gamma_1)$, helpfulness improves exactly when the new pinning point is closer to the peak than the old one, $|v^\star(s_{\text{disc}}) - v_{\text{pk}}| < |v^\star - v_{\text{pk}}|$. For $v^\star > v_{\text{pk}}$ this is equivalent to $v^\star(s_{\text{disc}}) > 2v_{\text{pk}} - v^\star$; the no-overshoot condition $v^\star(s_{\text{disc}}) \ge v_{\text{pk}}$, i.e. $s_{\text{disc}} \le w_1 b_1/\big(w_2 b_2 (1 + v_{\text{pk}})\big)$, is a simple sufficient condition (a mild overshoot still improves whenever $v^\star$ is far beyond the peak, but a large overshoot reduces helpfulness once $v^\star(s_{\text{disc}}) < 2v_{\text{pk}} - v^\star$). Because the gap is clipped and the margins are bounded sigmoid scores, $\hat{g}_1 \in [0,1]$ and hence $s_{\text{disc}} = 1 + \eta \hat{g}_1 \in [1, 1+\eta]$; GAPOR is therefore guaranteed non-overshooting whenever

$$\eta \;\le\; \frac{w_1 b_1}{w_2 b_2 \big(1 + \tilde{b}_1/(2\gamma_1)\big)} - 1, \tag{20}$$

i.e. the no-overshoot bound holds uniformly over all gaps for any gain below (20). Outside this gain range the discount can overshoot the peak, so the guarantee is a bounded-gain statement parameterized by η, r, and v_{pk}. This bound holds rigorously for reward-space reliability weighting (RLHF/GRPO); for joint DPO on a shared channel, temperature scaling is not algebraically equivalent to reward discounting (Proposition A.22(iii)), so the bound is a local approximation there. Its numerical value is instantiated over the identified parameter family in Table 16 (Remark A.8).

Remark A.8 (The shared-channel sweep form, and why it cannot identify the peak)**.** Treating β as the KL price of the binding regime, the first-order condition $w_1 b_1/(1+v) - w_2 b_2 = \beta v$ solves in closed form:

$$v(s) = \frac{-(1+s) + \sqrt{(1+s)^2 + 4s\,(r-1)}}{2s}, \qquad s = \beta/(w_2 b_2), \quad r = w_1 b_1/(w_2 b_2), \tag{21}$$

with $v(s) \to r - 1 = v^\star$ as $\beta \to 0$. The underlying helpfulness gold along the sweep is

$$g_1(\beta) = \alpha + \tilde{b}_1\, v(s) - \gamma_1\, v(s)^2, \tag{22}$$

whose peak lies at $v_{\text{pk}} = \tilde{b}_1/(2\gamma_1)$, i.e. at the temperature $\beta_{\text{peak}} = w_2 b_2 \big(\frac{r}{1+p} - 1\big)/p$ with $p = \tilde{b}_1/(2\gamma_1)$. **Identification caveat.** Eq. (22) is a model of the unbounded (logit-scale) gold; the observed held-out margin is a mean sigmoid score in $[0,1]$, a bounded monotone transform of g_1, so the quadratic form is a local approximation that is valid near the peak but goes negative for large v, outside the $[0,1]$ range of the observed margins (which stay in $[0.36, 0.45]$ over the whole sweep). The extended sweep of Protocol P7 supplies the rising limb (the gold peak lies at $\beta \approx 20$), and the fit is reported in Section 4.3 and Table 16: a least-squares fit of Eq. (22) to the 15 gold-sweep points reproduces the hump ($R^2 \approx 0.93$, fitted peak at $\beta \approx 20$) and confirms the criterion structurally ($v^\star/v_{\text{peak}} \approx 2.3 > 1$, and > 2 for net damage). The individual parameters $(\hat{r}, \tilde{b}_1, \gamma_1, w_2 b_2)$ are identified only up to a one-parameter scale ambiguity from the gold curve alone: the data fix the peak excess $\tilde{b}_1^2/(4\gamma_1) \approx 0.77$ and the ratio $v^\star/v_{\text{peak}} \approx 2.3$, while $\hat{r}$ and $w_2 b_2$ trade off through the horizontal scale. The non-overshoot bound of Eq. (20) is therefore reported as a range over the identified family (Table 16) rather than a single number.

Lemma A.9 (Budget–temperature correspondence, regime-corrected)**.** In the Gaussian policy model of Section 2, let D_i be the KL budget objective i spends at temperature β_i (the optimum $\pi^\star \propto \pi_{\text{ref}} \exp(r_i/\beta_i)$). Then (i) for a locally linear reward with gradient g_i, $D_i = w_i^2 \|g_i\|^2/(2\beta_i^2) \propto \beta_i^{-2}$; (ii) for the log proxy of (3), $v_i(1+v_i) = w_i b_i/\beta_i$ gives $D_i \approx w_i b_i/(2\beta_i) \propto \beta_i^{-1}$. The theory lives

in the large-budget regime, so the operative scaling is $D_i \propto \beta_i^{-1}$; the β^{-2} scaling (and the period-2 analysis of Proposition A.18) applies only in the weak-optimization regime, which Appendix B marks as misspecified. The correspondence is exact in the separable (orthogonal-channel) model, where objective i's spurious direction decouples from the others; on a shared channel the budget response of Algorithm 1 is analyzed directly in Proposition 2.9.

Full proof of Lemma A.9. The KL-regularized problem of objective i is $\max_\mu w_i p_i(\mu) - \beta_i \mathrm{KL}(\pi_\mu \| \pi_{\mathrm{ref}})$; its optimum satisfies $\pi^\star \propto \pi_{\mathrm{ref}} \exp(w_i p_i/\beta_i)$ (Rafailov et al., 2024), so w_i enters the reward, not the KL coefficient. (i) In the locally quadratic family the policy is Gaussian $\pi_\mu = \mathcal{N}(\mu, I)$ around the reference $\mathcal{N}(0, I)$; for a locally linear reward with gradient g_i, $\mu_i^\star = w_i g_i/\beta_i$ and $\mathrm{KL}(\pi_{\mu^\star} \| \pi_0) = \frac{1}{2} w_i^2 \|g_i\|^2/\beta_i^2 \propto \beta_i^{-2}$. (ii) For the log proxy, the first-order condition along v_i is $w_i b_i/(1+v_i) = \beta_i v_i$, i.e. $v_i(1+v_i) = w_i b_i/\beta_i$; in the large-budget regime $v_i \approx \sqrt{w_i b_i/\beta_i}$ and $D_i = v_i^2/2 \approx w_i b_i/(2\beta_i) \propto \beta_i^{-1}$, consistent with Lemma 2.2 (its multiplier satisfies $\lambda \approx \sum_j w_j b_j/(2\tilde{D})$, with β playing the role of λ). The linear-reward scaling of (i) is exactly the regime Appendix B identifies as misspecified for preference optimization: the operative scaling is (ii). □

Proposition A.10 (Quasi-static tracking of the spurious coordinates)**.** Fix objective i and temperature β_i, with the gold caps held saturated. The spurious-coordinate subproblem of the KL-regularized reward objective is

$$\max_{v \geq 0} \; f_i(v) \;=\; w_i b_i \log(1+v) \;-\; \tfrac{\beta_i}{2} v^2, \tag{23}$$

whose maximizer $v^\star$ satisfies $v^\star(1+v^\star) = w_i b_i/\beta_i$. Then f_i is $\mu = \beta_i$-strongly concave and $L = (w_i b_i + \beta_i)$-smooth on $v \geq 0$, and gradient ascent $v^{(t+1)} = v^{(t)} + \frac{1}{L} f_i'(v^{(t)})$ satisfies

$$\|v^{(t)} - v^\star\|^2 \;\leq\; \left(1 - \tfrac{\beta_i}{w_i b_i + \beta_i}\right)^t \|v^{(0)} - v^\star\|^2. \tag{24}$$

Consequently the spurious coordinate—and hence the damage $\Delta_i(v) = \gamma_i v^2 - \tilde{b}_i v$, which is Lipschitz on any bounded interval—tracks the quasi-static optimum with error decaying geometrically in the number T of updates between refreshes.

Proof. $f_i'(v) = w_i b_i/(1+v) - \beta_i v$ and $f_i''(v) = -w_i b_i/(1+v)^2 - \beta_i$; on $v \geq 0$ one has $-L \leq f_i''(v) \leq -\mu$ with $\mu = \beta_i$ and $L = w_i b_i + \beta_i$. The bound (24) is the standard gradient-ascent contraction for an L-smooth, μ-strongly-concave objective (the mirror of gradient descent on a strongly convex function). Because the damage $\Delta_i(v) = \gamma_i v^2 - \tilde{b}_i v$ is Lipschitz on any bounded interval, the same geometric decay transfers to $|\Delta_i(v^{(T)}) - \Delta_i(v^\star)|$; the refresh map of Proposition A.16 is exactly the $T \to \infty$ limit. □

Proposition A.11 (DPO gradient dynamics track the quasi-static optimizer)**.** In the separable (orthogonal-channel) model, fix the temperatures β_i and consider the joint population DPO loss along the spurious coordinates,

$$\mathcal{L}(v) \;=\; \sum_i w_i \Big[-q_i \log \sigma(\beta_i v_i) \;-\; (1-q_i) \log \sigma(-\beta_i v_i) \Big], \tag{25}$$

where q_i is objective i's preference probability and v_i its spurious coordinate. Then: (i) each term is strictly convex and $\beta_i^2/4$-smooth in v_i, since $0 < \partial^2 \mathcal{L}_i/\partial v_i^2 = \beta_i^2 \sigma(\beta_i v_i)\sigma(-\beta_i v_i) \leq \beta_i^2/4$, with unique minimizer $v_i^\star = \mathrm{logit}(q_i)/\beta_i$—by Rafailov's reparameterization this is the maximizer of the KL-regularized reward objective with the linear reward difference $\mathrm{logit}(q_i)$ (Proposition A.22(i)), i.e. the β_i^{-2} (weak-optimization) regime of Lemma A.9(i); it is the first-order linearization of the log-proxy maximizer $v^\star(1+v^\star) = w_i b_i/\beta_i$ (the β_i^{-1} large-budget regime of Lemma A.9(ii) on which the defense and convergence below are stated), so the two fixed points coincide only to first order; (ii) gradient descent with step size $\leq 4/\beta_i^2$ converges to $v_i^\star$ from any initialization; (iii) near $v_i^\star$ the loss is locally strongly convex with curvature $\beta_i^2 q_i(1-q_i)$, so convergence is linear once the iterate enters the region where $\sigma(\beta_i v_i)$ is bounded away from $\{0, 1\}$. Hence T gradient steps on the DPO loss converge, with a linear finite-T error once in the linear regime. The non-separable case is covered by Proposition A.12: the joint loss is still convex and gradient descent converges, but to a different fixed point (not the reward-model optimum, Proposition A.22(iii)).

Proof. (i) $\partial\mathcal{L}_i/\partial v_i = \beta_i(\sigma(\beta_i v_i) - q_i)$ and $\partial^2\mathcal{L}_i/\partial v_i^2 = \beta_i^2\sigma(\beta_i v_i)\sigma(-\beta_i v_i)$, which is strictly positive and bounded by $\beta_i^2/4$; the minimizer solves $\sigma(\beta_i v_i^\star) = q_i$, i.e. $v_i^\star = \mathrm{logit}(q_i)/\beta_i$. (ii) is standard gradient-descent convergence for a convex, smooth objective; (iii) at $v_i^\star$ the curvature is $\beta_i^2 q_i(1-q_i) > 0$, and by continuity $\partial^2\mathcal{L}_i/\partial v_i^2 \geq \frac{1}{2}\beta_i^2 q_i(1-q_i)$ in a neighborhood, giving local strong convexity and linear convergence there. □

Proposition A.12 (Joint DPO dynamics converge on any shared channel)**.** Let objectives $i \in [K]$ share one direction v with signs $s_i \in \{\pm 1\}$ (objective i's log-ratio along v is $z_i = s_i v$). The joint population DPO loss

$$\mathcal{L}(v) \;=\; \sum_i w_i\Big[-q_i \log\sigma(s_i\beta_i v) \;-\; (1-q_i)\log\sigma(-s_i\beta_i v)\Big], \tag{26}$$

is strictly convex, with Hessian $\mathcal{L}''(v) = \sum_i w_i\beta_i^2\,\sigma(s_i\beta_i v)\,\sigma(-s_i\beta_i v) \in \big(0,\ \frac{1}{4}\sum_i w_i\beta_i^2\big]$, so gradient descent with step size $\leq 4/\sum_i w_i\beta_i^2$ converges to the unique minimizer from any initialization, and linearly once the iterate is near the minimizer (where the curvature $\mathcal{L}''(v^\star) > 0$). The minimizer is characterized by the stationarity condition

$$\sum_i w_i\, s_i\, \beta_i\, \big(\sigma(s_i\beta_i v) - q_i\big) \;=\; 0. \tag{27}$$

In the separable (orthogonal) case this reduces to the per-objective DPO fixed point of Proposition A.11 (the β_i^{-2} linear-reward regime of Lemma A.9(i)), which coincides with the reward-model log-proxy optimizer only to first order; in the non-separable case it is generally a different fixed point—not the reward-discount optimum (Proposition A.22(iii))—but the dynamics still converge to it.

Proof. Each term $-q\log\sigma(s\beta v)$ and $-(1-q)\log\sigma(-s\beta v)$ is convex in v (the logistic loss composed with a linear map), with second derivative $\beta^2\sigma(s\beta v)\sigma(-s\beta v) > 0$; summing gives the Hessian $\mathcal{L}''(v)$ above, which is strictly positive and bounded by $\frac{1}{4}\sum_i w_i\beta_i^2$. The stationarity condition $\mathcal{L}'(v) = 0$ is (27). Linear convergence near the minimizer follows from $\mathcal{L}''(v^\star) > 0$ and continuity. □

Remark A.13 (Which budget–temperature scaling does DPO instantiate?)**.** Lemma A.9 gives two scalings: $D_i \propto \beta_i^{-2}$ for a linear reward (the weak-optimization regime) and $D_i \propto \beta_i^{-1}$ for the log proxy (the large-budget regime). These correspond to two different objectives: the reward model (Assumption 2.1) maximizes the log proxy, while DPO's implicit reward is linear in the log-ratio (Proposition A.11), so DPO instantiates the β_i^{-2} regime. The defense of Proposition 3.1 and the convergence map of Proposition A.16 are stated in the β_i^{-1} (reward-model) regime, which is the operative one for RLHF-style reward maximization (Assumption 2.1); the DPO experiments of Section 4, however, instantiate the β_i^{-2} regime, under which the shrinkage is squared, $\Delta_i \leftarrow \Delta_i/(1+\eta\hat{g}_i)^2$, and the loop map is $f(x) = A/(1+\eta\kappa x)^2$, which admits a period-2 cycle for large gain (Proposition A.18). The empirical gain sweep (no oscillation up to $\eta = 2$, Table 20) is consistent with operating below the oscillation threshold, but the β_i^{-1} no-oscillation guarantee does not transfer verbatim to DPO.

Full proof of Proposition 3.1. In the large-budget regime $\Delta_i(D_i) \approx 2\gamma_i D_i$ (Theorem 2.3, with $D_i = c_i D$ the per-objective budget share). Substituting $D_i = c_i D/(1+\eta\hat{g}_i)$ gives $\Delta_i^{\mathrm{GAPOR}}(D) \approx 2\gamma_i c_i D/(1+\eta\hat{g}_i) = \Delta_i^{\mathrm{naive}}(D)/(1+\eta\hat{g}_i)$. For the collapse point, the rescaled gold trajectory is $G_i(D) = a_i C_i + \tilde{b}_i\sqrt{2c_i D/(1+\eta\hat{g}_i)} - 2\gamma_i c_i D/(1+\eta\hat{g}_i)$; differentiating in the shared budget D, $\partial G_i/\partial D = \frac{\tilde{b}_i\sqrt{2c_i/(1+\eta\hat{g}_i)}}{2\sqrt{D}} - \frac{2\gamma_i c_i}{1+\eta\hat{g}_i}$, which vanishes exactly at $D = \tilde{b}_i^2(1+\eta\hat{g}_i)/(8\gamma_i^2 c_i) = D_i^\star(1+\eta\hat{g}_i)$. □

Proposition A.14 (Balanced fixed point)**.** Suppose the gap tracks damage linearly with a common slope, $\hat{g}_i = \kappa\Delta_i$ with $\kappa > 0$, and the refresh dynamics reach a fixed point. Then each objective satisfies the fixed-point equation

$$\Delta_i \;=\; \frac{2\gamma_i c_i D}{1+\eta\kappa\Delta_i}, \tag{28}$$

whose positive solution is unique ($x \mapsto x(1+\eta\kappa x)$ is strictly increasing on $x \geq 0$), and the allocation ratio is

$$\frac{\Delta_i}{\Delta_j} = \frac{\gamma_i c_i}{\gamma_j c_j} \cdot \frac{1+\eta\kappa\Delta_j}{1+\eta\kappa\Delta_i} < \frac{\gamma_i c_i}{\gamma_j c_j} \quad \text{whenever } \Delta_i > \Delta_j. \tag{29}$$

That is, under the common-slope tracking model, GAPOR strictly reduces the allocation ratio below its naive value $\gamma_i c_i / \gamma_j c_j$ (toward one, without claiming full equalization): no objective monopolizes the overoptimization budget, and the differential effect of Section 2 is attenuated.

Remark A.15 (The common-slope assumption is not derived)**.** The exact linear model $\hat{g}_i = \kappa\Delta_i$ with objective-independent κ is an assumption of Proposition A.14, not a consequence of the bridge. In general the gap–damage slope is objective-specific (Remark A.3), and two increasing gap functions $\hat{g}_1(d) = d$ and $\hat{g}_2(d) = 10d$ already reverse the correction's sign at some (d_1, d_2): the rebalancing in Eq. (29) is guaranteed only when the tracking slopes are comparable across objectives. It also holds only while the gap is far from saturation: the experiment's margins are bounded sigmoid scores in $[0, 1]$, so the clipped gap satisfies $0 \leq \hat{g}_i/\Delta_i \leq 1/\Delta_i \to 0$ as $\Delta_i \to \infty$, and the linear tracking $\hat{g}_i = \kappa\Delta_i$ is a local description valid for moderate Δ_i before the sigmoid saturates.

Full proof of Proposition A.14. At a fixed point $\hat{g}_i = \kappa\Delta_i$, so $\beta_i = \beta(1+\eta\kappa\Delta_i)$ and each objective satisfies $\Delta_i = 2\gamma_i c_i D/(1+\eta\kappa\Delta_i)$; the map $x \mapsto x(1+\eta\kappa x)$ is strictly increasing on $x \geq 0$, so the positive fixed point is unique. Dividing the two fixed-point equations gives (29), and since $(1+\eta\kappa\Delta_j)/(1+\eta\kappa\Delta_i) < 1$ whenever $\Delta_i > \Delta_j$, the allocation ratio is strictly smaller than $\gamma_i c_i/\gamma_j c_j$: the damage is rebalanced toward uniformity. □

Proposition A.16 (Convergence of the closed loop)**.** Fix objective i and suppose the loop (quasi-static refresh, common-slope tracking of Proposition A.14, memorization subdominant) refreshes with the synchronous map

$$\Delta^{(t+1)} = \frac{A}{1+\eta\kappa\,\Delta^{(t)}}, \qquad A = 2\gamma_i c_i D, \tag{30}$$

with $\kappa > 0$ the tracking slope of Proposition A.14. Let $\Delta^\star$ be the unique positive fixed point, $\Delta^\star(1+\eta\kappa\Delta^\star) = A$. Then, for this map:

(i) (Contraction) f maps $[0, A]$ into itself with $\sup_{[0,A]} |f'| = A\eta\kappa$; if $A\eta\kappa < 1$, f is a contraction and $\Delta^{(t)} \to \Delta^\star$ from any $\Delta^{(0)} \geq 0$ linearly;

(ii) (Global convergence of the map) for every $\eta\kappa \geq 0$, $\Delta^{(t)} \to \Delta^\star$ from any $\Delta^{(0)} \geq 0$: the second iterate $g = f \circ f$ is increasing and $g(x) - x$ has the sign of $A - x - \eta\kappa x^2$, so both subsequences converge monotonically to $\Delta^\star$;

(iii) (Regime dependence of the linear-reward map) in the weak-optimization regime ($D_i \propto \beta_i^{-2}$) the corresponding map $f(x) = A/(1+\eta\kappa x)^2$ admits a period-2 cycle when $\eta\kappa\Delta^\star > 1$ (equivalently $A\eta\kappa > 4$; the full statement is Proposition A.18). The diminishing-returns map of Eq. (30) has no such cycle.

The fixed point is exactly the balanced value of Proposition A.14. These statements are properties of the quasi-static map: the per-objective loops are assumed to decouple to first order, and Eq. (29) is the joint consequence of the per-objective fixed points. Finite-T gradient updates track the per-objective optima in the reward model (geometrically, Proposition A.10), on the separable DPO loss (Proposition A.11), and on any shared channel the joint DPO loss is convex so gradient descent converges (Proposition A.12); the exact nonlinear loop need not be globally stable (Remark A.17).

Full proof of Proposition A.16. The map $f(x) = A/(1+\eta\kappa x)$ is decreasing with $f'(x) = -A\eta\kappa/(1+\eta\kappa x)^2$, so $\sup_{[0,A]} |f'| = A\eta\kappa$, and f maps $[0, A]$ into itself ($f(0) = A$, $f(A) < A$). (i) If $A\eta\kappa < 1$, f is a contraction and Banach's theorem gives convergence to the unique fixed point from any $\Delta^{(0)} \geq 0$ at a linear rate. (ii) The second iterate is $g(x) = f(f(x)) = A(1+\eta\kappa x)/(1+\eta\kappa x + \eta\kappa A)$; a direct computation gives $g(x) - x = (A - x - \eta\kappa x^2)/(1+\eta\kappa x + \eta\kappa A)$, whose numerator is decreasing on $x \geq 0$ and vanishes exactly at $x = \Delta^\star$ (since $\Delta^\star(1+\eta\kappa\Delta^\star) = A$). Hence $g(x) > x$ for $x < \Delta^\star$ and $g(x) < x$ for $x > \Delta^\star$; g is increasing and maps $[0, A]$ into itself, so its iterates converge to $\Delta^\star$, and the odd subsequence $f(g^{(t)}(x))$ converges to $f(\Delta^\star) = \Delta^\star$. (iii) See Proposition A.18. □

Remark A.17 (The linear-regime map and its period-2 cycle)**.** In the weak-optimization (locally linear) regime, Lemma A.9(i) gives $D_i \propto \beta_i^{-2}$, so (15) is replaced by $D_i = D/(1+\eta\hat{g}_i)^2$ and the loop map becomes $f(x) = A/(1+\eta\kappa x)^2$. Here $|f'(\Delta^\star)| = 2\eta\kappa\Delta^\star/(1+\eta\kappa\Delta^\star)$, so the fixed point repels when $\eta\kappa\Delta^\star > 1$; the second iterate $g = f \circ f$ is increasing with $g(0) = f(A) > 0$ and $g(A) = f(f(A)) < A$; writing $h(x) = g(x) - x$, $h'(\Delta^\star) = f'(\Delta^\star)^2 - 1 > 0$, so h is negative just below $\Delta^\star$ and positive just above it; since $h(0) > 0$ and $h(A) < 0$, by continuity h crosses zero at least once in $(0, \Delta^\star)$ and at least once in $(\Delta^\star, A)$, each such $y \neq \Delta^\star$ yielding a period-2 orbit $\{y, f(y)\}$. This is the quantitative form of over-correction oscillation for large η in the linear-reward regime. The complete statement, with the exact contraction, global-convergence, and bifurcation thresholds, is Proposition A.18 below. Caution: the diminishing-returns map $f(x) = A/(1+\eta\kappa x)$ of Eq. (30) has no period-2 cycle (Proposition A.16), but the exact nonlinear loop does not inherit this. With the log proxy $p(v) = \log(1+v)$ and gold $g(v) = g_0 + v - v^2$ (so Condition 2.6 holds exactly), the measured gap is $S(v) = \log(1+v) - v + v^2 = \Delta(v) + o(\Delta(v))$ and the refresh map $F(x) = S(v(x))$ with $v(1+v) = 1/\big(\beta(1+x)\big)$ satisfies $F'(x) = -v(x)^2/(1+x)$ (GAPOR's $\beta_i = \beta(1+x)$ is the KL coefficient, so a larger gap increases β_i and decreases v). Choosing $\beta^{-1} = V(1+V)[1+S(V)]$ for any $V \geq 3$ gives a fixed point $x^\star = S(V)$ with $v(x^\star) = V$ and $|F'(x^\star)| = V^2/(V^2 - V + 1 + \log(1+V)) > 1$: the equilibrium is repelling at arbitrarily large coordinates (numerically the map also exhibits a two-cycle). The "unconditional" convergence of Proposition A.16 is therefore a property of the map, not of the exact loop.

Proposition A.18 (Bounded-gain convergence of the DPO (β_i^{-2}) loop)**.** Consider the per-objective loop in the weak-optimization (linear-reward) regime of Lemma A.9(i), where $D_i \propto \beta_i^{-2}$ and the synchronous refresh map of Remark A.13 reads

$$\Delta^{(t+1)} \;=\; f(\Delta^{(t)}) \;=\; \frac{A}{(1+\eta\kappa\,\Delta^{(t)})^2}, \qquad A = 2\gamma_i c_i D > 0, \qquad \eta\kappa \geq 0, \tag{31}$$

with unique positive fixed point $\Delta^\star$, $\Delta^\star(1+\eta\kappa\Delta^\star)^2 = A$. Then:

(i) (**Self-map and monotone second iterate**) f is strictly decreasing and maps $[0, A]$ into itself; $g = f \circ f$ is strictly increasing and maps $[0, A]$ into itself, so every orbit of g converges monotonically to a fixed point of g.

(ii) (**Local stability**) $|f'(\Delta^\star)| = \frac{2\eta\kappa\Delta^\star}{1+\eta\kappa\Delta^\star}$, and $\eta\kappa\Delta^\star \lessgtr 1 \iff A\eta\kappa \lessgtr 4$; hence $\Delta^\star$ is attracting iff $A\eta\kappa < 4$ and repelling iff $A\eta\kappa > 4$.

(iii) (**Contraction**) $\sup_{[0,A]} |f'| = |f'(0)| = 2A\eta\kappa$; if $2A\eta\kappa < 1$ then f is a contraction and $\Delta^{(t)} \to \Delta^\star$ linearly from any $\Delta^{(0)} \geq 0$.

(iv) (**Global convergence**) Writing $y = 1 + \eta\kappa\Delta$, $g'(\Delta) = \frac{4A^2\eta^2\kappa^2\,y^3}{(y^2 + A\eta\kappa)^3}$ with $\max_{\Delta\geq 0} g'(\Delta) = \frac{\sqrt{A\eta\kappa}}{2}$. If $A\eta\kappa < 4$ then $g' < 1$ everywhere, $h(\Delta) := g(\Delta) - \Delta$ is strictly decreasing with $h(0) > 0 > h(A)$ and unique root $\Delta^\star$; hence $g(\Delta) > \Delta$ for $\Delta < \Delta^\star$ and $g(\Delta) < \Delta$ for $\Delta > \Delta^\star$, and $\Delta^{(t)} \to \Delta^\star$ from any $\Delta^{(0)} \geq 0$.

(v) (**Period-2 bifurcation**) If $A\eta\kappa > 4$ then $|f'(\Delta^\star)| > 1$ and $h'(\Delta^\star) = f'(\Delta^\star)^2 - 1 > 0$; since $h(0) > 0$ and $h(A) < 0$, h has a root $\Delta_1 \in (0, \Delta^\star)$ and a root $\Delta_2 \in (\Delta^\star, A)$ with $f(\Delta_1) = \Delta_2$ and $f(\Delta_2) = \Delta_1$: a period-2 (over-correction) orbit. The map undergoes a period-doubling bifurcation at $A\eta\kappa = 4$.

The β_i^{-2} loop therefore differs sharply from the β_i^{-1} loop of Proposition A.16: the latter converges for every gain $\eta\kappa \geq 0$ with no cycle, whereas the former is stable only below the threshold $A\eta\kappa < 4$ and over-corrects (oscillates) beyond it—the bounded-gain guarantee of Remark A.13 made precise.

Proof. $f'(\Delta) = -2A\eta\kappa/(1+\eta\kappa\Delta)^3 < 0$, $f(0) = A$, and $f(A) = A/(1+\eta\kappa A)^2 < A$ with $0 < f(\Delta) \leq A$ on $[0, A]$, giving (i); $g = f \circ f$ is increasing as the self-composition of a decreasing map. The fixed point is unique because $\Delta \mapsto \Delta(1+\eta\kappa\Delta)^2$ is strictly increasing on $\Delta \geq 0$ (derivative $(1+\eta\kappa\Delta)(1+3\eta\kappa\Delta) > 0$).

(ii) $|f'(\Delta^\star)| = \frac{2A\eta\kappa}{(1+\eta\kappa\Delta^\star)^3} = \frac{2\eta\kappa\Delta^\star(1+\eta\kappa\Delta^\star)^2}{(1+\eta\kappa\Delta^\star)^3} = \frac{2\eta\kappa\Delta^\star}{1+\eta\kappa\Delta^\star}$, using $A = \Delta^\star(1+\eta\kappa\Delta^\star)^2$. Writing $t = \eta\kappa\Delta^\star$, $A\eta\kappa = t(1+t)^2$ is strictly increasing in t, so $t \lessgtr 1 \iff A\eta\kappa \lessgtr 4$.

(iii) $|f'(\Delta)| = 2A\eta\kappa/(1+\eta\kappa\Delta)^3$ is maximized at $\Delta = 0$.

(iv) With $y = 1+\eta\kappa\Delta$, $g'(\Delta) = f'(f(\Delta))\, f'(\Delta) = \frac{2A\eta\kappa}{(1+\eta\kappa f(\Delta))^3}\frac{2A\eta\kappa}{(1+\eta\kappa\Delta)^3} = \frac{4A^2\eta^2\kappa^2}{(1+A\eta\kappa/y^2)^3\, y^3} = \frac{4A^2\eta^2\kappa^2\, y^3}{(y^2+A\eta\kappa)^3}$. Writing $\varphi(y) = y/(y^2+A\eta\kappa)$, $g' = 4A^2\eta^2\kappa^2\,\varphi(y)^3$; φ is maximized at $y = \sqrt{A\eta\kappa}$ with value $1/(2\sqrt{A\eta\kappa})$, so $\max_{\Delta\geq 0} g'(\Delta) = \frac{4A^2\eta^2\kappa^2}{8(A\eta\kappa)^{3/2}} = \frac{\sqrt{A\eta\kappa}}{2}$. If $A\eta\kappa < 4$ then $g' < 1$ everywhere, so $h' = g' - 1 < 0$ and h is strictly decreasing; $h(0) = f(f(0)) = f(A) > 0$ and $h(A) = f(f(A)) - A < 0$, so h has exactly one root, which is $\Delta^\star$ since $h(\Delta^\star) = f(\Delta^\star) - \Delta^\star = 0$. Thus $h(\Delta) > 0$ on $[0, \Delta^\star)$ and $h(\Delta) < 0$ on $(\Delta^\star, A]$, i.e., $g(\Delta) > \Delta$ for $\Delta < \Delta^\star$ and $g(\Delta) < \Delta$ for $\Delta > \Delta^\star$; the monotone iterates of g converge to the unique fixed point $\Delta^\star$, and the odd subsequence $f(g^{(t)}(\Delta^{(0)}))$ converges to $f(\Delta^\star) = \Delta^\star$ by continuity.

(v) If $A\eta\kappa > 4$ then $\eta\kappa\Delta^\star > 1$ and $h'(\Delta^\star) = f'(\Delta^\star)^2 - 1 = \left(\frac{2\eta\kappa\Delta^\star}{1+\eta\kappa\Delta^\star}\right)^2 - 1 > 0$. Since $h(0) > 0$ and h is negative immediately to the left of $\Delta^\star$ ($h(\Delta^\star) = 0$ with $h'(\Delta^\star) > 0$), continuity gives a root $\Delta_1 \in (0, \Delta^\star)$; likewise h is positive immediately to the right of $\Delta^\star$ while $h(A) < 0$, giving a root $\Delta_2 \in (\Delta^\star, A)$. For Δ_1, $g(\Delta_1) = \Delta_1$ means $f(f(\Delta_1)) = \Delta_1$; setting $\Delta_2 = f(\Delta_1)$ gives $f(\Delta_2) = f(f(\Delta_1)) = \Delta_1$, and $\Delta_1 \neq \Delta^\star$ implies $\Delta_2 \neq \Delta^\star$, so $\{\Delta_1, \Delta_2\}$ is a genuine period-2 orbit. □

Remark A.19 (Theory: Regime, and the visible part of the hump)**.** The trajectory form of (7) holds once the gold caps saturate, i.e., for budgets D large enough that every objective has exhausted its finite task performance; this is the regime assumed by the statements below. Whether the rising limb of the hump is observable depends on whether saturation precedes the peak; in our benchmark it is observable (Protocol P7).

Remark A.20 (Theory: The role of the user's weights)**.** The effective hackability c_i couples the user's weight w_i with the proxy's weakness b_i. Equal weights recover the pure "weakest proxy" effect; extreme weights amplify it (Proposition A.1). What reweighting cannot do is shrink damage without moving the user's operating point (Remark 2.5); a remedy must therefore act on the budget, not the weights (Section 3).

Remark A.21 (Method: Small-gain scaling)**.** The condition $A\eta\kappa < 1$ is the textbook small-gain condition of feedback control: the permissible gain η scales inversely with the naive overoptimization A and the tracking slope κ, the more damage there is to correct, the more conservative the controller must be. Unlike the linear-regime map, the large-budget map remains convergent beyond this threshold (Proposition A.16(ii)); the condition governs the convergence rate, not the outcome. (For the DPO/linear-reward map of Proposition A.18 the corresponding thresholds are $2A\eta\kappa < 1$ for contraction and $A\eta\kappa < 4$ for stability, so the "unconditional" reading of the small-gain condition is specific to the reward-model β_i^{-1} map.)

Finally, the same one-line change has a reward-space reading whose scope must be stated precisely, because the equivalence holds off the shared channel but not on it.

Proposition A.22 (Method: reward-space reading, with exact scope)**. (i) Single objective.** For objective i alone, DPO with temperature β_i is equivalent in the population limit to KL-regularized reward maximization with reward r_i and KL coefficient β_i, whose optimum is $\pi^\star \propto \pi_{\text{ref}} \exp(r_i/\beta_i)$ (Rafailov et al., 2024); hence $\beta_i \leftarrow \beta(1+\eta\hat{g}_i)$ is equivalent to $r_i \leftarrow r_i/(1+\eta\hat{g}_i)$ at fixed β. **(ii) Separable joint model.** In the orthogonal-channel model of Section 2 the per-objective losses decouple along the orthogonal spurious directions, so the joint loss $\sum_i w_i \mathcal{L}_i(\theta; \beta_i)$ inherits the same per-direction reading: $\beta_i \leftarrow \beta(1+\eta\hat{g}_i)$ discounts objective i's spurious reward by $(1+\eta\hat{g}_i)^{-1}$. **(iii) Non-separable case, precise.** When objectives share a direction (Section 2.3) or their preference directions are correlated, the joint DPO loss is not a scalarized-reward objective. In the population limit, with per-objective preference probabilities q_i, orientation signs $s_i \in \{\pm 1\}$ (objective 2 of Section 2.3 has $s_2 = -1$), and a shared coordinate z, the joint stationarity condition is

$$\sum_i w_i\, s_i\, \beta_i \left(\sigma(s_i\beta_i z) - q_i\right) \;=\; 0, \tag{32}$$

whereas the discounted scalarized-reward optimum at a shared β is $z^\star = \sum_i w_i \operatorname{logit}(q_i)/\big(\beta(1+\eta\hat{g}_i)\big)$. Scaling β_i enters (32) both as a linear prefactor on objective i's gradient contribution and inside the sigmoid argument, so it can shift the joint stationary point toward objective i's preference rather than suppressing it. These two conditions coincide only in the degenerate single-objective

limit ($q_i = q$ for all i with uniform discount and common sign) or on a decoupled (separable) channel. Off the separable channel, per-objective temperature scaling and per-objective reward discounting are therefore different interventions; the reward-space form is a distinct, principled intervention (reliability weighting), not an algebraic identity.

Proof. (i) is the standard DPO reparameterization: the population minimizer of the DPO loss with temperature β_i equals the maximizer of $\mathbb{E}[r_i] - \beta_i \,\mathrm{KL}(\pi \| \pi_{\mathrm{ref}})$ (Rafailov et al., 2024). Writing $\beta_i = \beta(1 + \eta\hat{g}_i)$ yields $\pi_{\mathrm{ref}} \exp\big((r_i/(1 + \eta\hat{g}_i))/\beta\big)$. (ii) In the orthogonal model the spurious direction v_i appears only in objective i's loss, so the joint first-order condition along v_i coincides with the single-objective one. (iii) For a shared coordinate z with a uniform reference and orientation signs $s_i \in \{\pm 1\}$, objective i's population loss is $\mathcal{L}_i = -q_i \log \sigma(s_i\beta_i z) - (1 - q_i) \log \sigma(-s_i\beta_i z)$, whose derivative is $s_i\beta_i(\sigma(s_i\beta_i z) - q_i)$; summing over i gives (32). The discounted-reward optimum is the minimizer of $-\sum_i w_i r_i/(1 + \eta\hat{g}_i) + \beta \,\mathrm{KL}$, i.e. $z^\star$ above. The two coincide only in the degenerate single-objective limit ($q_i = q$ for all i, uniform discount, and common sign, so $\beta_i = \beta$ and $s_i = s$ for all i; then (32) reduces to $\sigma(s\beta z) = q$ and $z^\star = \mathrm{logit}(q)/(s\beta)$), or when the directions decouple as in (ii). The one-pair counterexample shows the gap concretely: with equal weights, a uniform reference, and $q_1 = 3/4$, $q_2 = 1/4$, the joint DPO loss with $\beta_1 = 2\beta$, $\beta_2 = \beta$ has a minimizer favoring objective 1's preferred response, whereas the KL-regularized weighted-reward objective with r_1 halved favors objective 2's. □

Proposition A.22 gives the method a second, intuitive reading: the more an objective's proxy diverges from its held-out signal, the less weight its reward carries in the total loss. This is an instance of reliability-weighted preference learning, which has been proposed on statistical grounds (Singha, 2026); here the discount factor is derived from the allocation law rather than postulated, which fixes both the signal ($\hat{g}_i$) and the functional form ($1/(1 + \eta\hat{g}_i)$).

B Misspecification Robustness of the Allocation Law

The model of Section 2 is deliberately minimal, so a skeptic may worry that the allocation law is an artifact of its functional forms (the logarithmic proxy, the quadratic fragility, the orthogonal spurious directions). We therefore re-run the verification under systematic misspecification: each assumption is perturbed while the others are held fixed, and the optimizer is solved exactly (the full first-order condition, never the asymptotic allocation) throughout. Each perturbed family is evaluated over 1000 independent random parameter draws with equal weights; the observed allocation order, the weakest-first collapse prediction, and the maximum allocation-share error are compared against the unperturbed model.

Table 5 reports two kinds of stability. (i) Qualitative stability. The allocation order and the weakest-first collapse prediction are structural: under proxy-form perturbation, observation noise on the perceived hackability ($\pm 30\%$), a cubic correction to the gold fragility, and correlated spurious directions ($\rho = 0.5$), the predicted order matches the observed order in 76–100% of draws (chance: 17%) and the weakest-first prediction in 85–92% (chance: 33%). (ii) Quantitative stability. The proportionality $\Delta_i \propto \gamma_i w_i b_i$ is the refinement of the diminishing-returns regime: for proxies with diminishing returns (log, saturating, sublinear) and for the noise and correlation perturbations, the maximum allocation error stays below 18.6%; it degrades only for the linear proxy (18.6%), while the superlinear proxy stays comparable (13.0%), where overoptimization concentrates even more strongly on the hackable objective. KL-regularized preference optimization lives exactly in the diminishing-returns regime, the DPO loss is a sigmoid of the log-ratio and the RLHF objective is log-sum-exp, so the quantitative form the theorem derives is the one relevant to practice, and the qualitative law is structural.

C Extended Related Work

Overoptimization and Goodhart effects. The single-objective overoptimization curve (gold quality first rising and then falling as the KL budget grows) was established by scaling-law studies (Gao et al., 2022); the formal treatment of gaming and Goodhart effects goes back to (Skalse et al., 2022; Manheim & Garrabrant, 2018; Krakovna et al., 2020). In the principal–agent view, unlimited optimization of a proxy that covers only part of the attributes can drive true utility arbitrarily

Table 5: Robustness of the allocation law to model misspecification (exact optimizer, 1000 random draws each, equal weights; chance levels are 17% for the order and 33% for weakest-first). The upper block keeps diminishing returns, the regime of preference optimization; the lower block removes it.

Perturbation	order match	weakest-first	max share err
None (log proxy, quadratic gold)	100%	91%	1.9%
Perceived hackability, ±30% noise	80%	87%	11.9%
Saturating proxy $v/(1+v)$	86%	85%	15.8%
Sublinear proxy $v^{0.7}$	79%	90%	12.5%
Cubic gold fragility	82%	90%	11.3%
Correlated directions (ρ=0.5)	89%	92%	7.6%
Linear proxy	80%	89%	18.6%
Superlinear proxy $v^{1.3}$	76%	86%	13.0%

low (Zhuang & Hadfield-Menell, 2020), and quantitative single-objective Goodhart analyses with provable gap-driven early stopping exist (Karwowski et al., 2023). For AI evaluation the principal–agent view was recently re-derived as a static per-dimension distortion index (Wang & Huang, 2026). Our allocation law is the dynamic counterpart of that static index: it describes how the distortion grows with optimization strength, where it lands among objectives, and when each objective collapses ($D_i^\star$). The length channel is the best-documented spurious feature: RLHF models exploit verbosity systematically (Singhal et al., 2023), and disentangling length from quality requires either explicit length regularization (Park et al., 2024) or length-controlled evaluation (Dubois et al., 2024; Chen et al., 2024). For the classical multi-objective optimization machinery (Pareto concepts, scalarization theory, Pareto-front learning) we follow the treatment of (Roijers et al., 2013).

The preference-optimization family. RLHF proceeds by reward modeling plus policy optimization (Christiano et al., 2017; Ziegler et al., 2020; Stiennon et al., 2022; Ouyang et al., 2022). DPO reparameterizes the optimal KL-regularized policy so that the temperature β becomes the single explicit budget knob (Rafailov et al., 2024); IPO removes the Bradley–Terry overfitting pathology (Azar et al., 2023); KTO relaxes the paired-data requirement (Ethayarajh et al., 2024); RSO adds rejection sampling (Liu et al., 2023); SimPO drops the reference model and normalizes by length with a target margin (Meng et al., 2024); length-regularized DPO attacks the verbosity channel directly (Park et al., 2024); and GRPO is the policy-gradient variant sharing the same KL-regularized template (Shao et al., 2024). What all of these share is a single global budget: one temperature for every objective. Instance-level or adaptive temperature schemes exist—f-DPO introduces instance-level divergence regularization (Wang et al., 2023) and sDPO adapts β during training (Kim et al., 2024)—but neither is per objective in the sense needed to track per-objective overoptimization, and we are not aware of prior work that drives the per-objective temperature from a measured proxy–gold gap.

Multi-objective alignment. Fine-grained RLHF aggregates per-dimension reward heads (Wu et al., 2023); Safe-RLHF couples reward maximization with a constraint value head (Dai et al., 2023); MODPO extends DPO to per-dimension margins (Zhou et al., 2024); Rewarded Soups interpolates single-objective experts into a frontier (Ramé et al., 2023) and WARP extends the idea to weight-averaged multi-reward policies (Ramé et al., 2024); Panacea learns a low-rank preference subspace (Zhong et al., 2024); AMoPO adapts per-dimension weights from the policy's own generation statistics (Liu et al., 2025); CPO conditions on the preference vector (Guo et al., 2024); and classical multi-gradient methods resolve inter-objective gradient conflicts (Yu et al., 2020; Liu et al., 2021) while non-convex scalarizations such as Tchebycheff or ϵ-constraint replace the weighted sum. Multi-task loss weighting with signal-driven per-objective coefficients is classical (Kendall et al., 2018; Chen et al., 2018). Li et al. (2025) most recently put multiple-gradient descent inside preference optimization: GAPO adaptively rescales each objective's gradient to balance conflicting ones, and its preference-weighted variant P-GAPO targets user-specific Pareto solutions. The relationship to our work is complementary and precise. GAPO answers "in which direction and at which trade-off should the policy move": it balances the gradients that compete within a step. It never measures the proxy–gold gap, so it has no signal with which to detect that the hackable ob-

jective is absorbing the damage; our experiments show this damage is unchanged under direction balancing (Table 1, GAPO gap $+0.174$ vs. DPO-LW $+0.174$), which we read as consistent with—not a causal proof of—the allocation view that the budget, not the direction of the update, governs overoptimization. GAPOR acts on the per-objective budget that GAPO and all other direction- or trade-off-level methods share but none adapts; and our weight-locus sweep (Section 4) is the empirical counterpart of P-GAPO's preference targeting, evaluated against the same frontier questions.

AMoPO in detail. Liu et al. (2025) model the per-token generation probabilities as per-dimension Gaussians, normalize them by softmax into per-dimension weights, and use no reward model, no reference model, and a fixed β. Its stated goal is dynamic balance across preference dimensions rather than overoptimization control; under our law, its actuator (the weights) moves the trade-off point without contracting any objective's budget (Remark 2.5), which is visible in our experiments as a large conciseness drop at a reduced helpfulness gap (Table 1). AMoPO runs in its reference-free form and constrained RLHF in its reward-space CMDP form (their original objectives, not ported); the DPO-based methods share the prompt-free, length-normalized scoring scaffold of Remark D.1.

Defenses against overoptimization. Early stopping and KL regularization are the canonical remedies (Gao et al., 2022; Prechelt, 1998); constrained RLHF formulates policy optimization as a constrained MDP in reward space, enforcing per-component improvement constraints with the KL term as a soft penalty (Moskovitz et al., 2023); we run it in this original form; reward-model ensembling mitigates hacking but does not eliminate it (Eisenstein et al., 2024); and uncertainty-based discounting scales the reward by model confidence (Singha, 2026). All of these act at a coarser level than the per-objective budget—one shared KL budget, cap, or temperature across objectives (constrained RLHF constrains each component's reward but keeps a single global KL), or they target one known channel (length). GAPOR is per-objective and signal-driven within the preference-optimization family, spending regularization exactly where the measured gap opens; its closest conceptual relative is gap-driven early stopping (Karwowski et al., 2023), which we isolate against in Protocol P3 (Appendix D).

Inference-time and robust multi-objective variants. Differential overoptimization also governs inference-time selection. Best-of-N re-ranking samples N completions per prompt and scores them with the (scalarized) proxy; because the re-ranker evaluates the same hackable proxy, a shared sample budget concentrates its Goodhart damage on the most hackable objective exactly as Theorem 2.3 describes—the same selection-scaling behavior Gao et al. (2022) studies for the single-objective case—and Pareto Best-of-N variants inherit this without a per-objective gap signal to temper it. Distributionally robust and min-max scalarizations (maximizing the worst-case objective) likewise do not resolve the problem: the worst-case objective is computed against the proxy, so when that proxy is itself hackable the min-max optimum still drives the hackable channel's damage; robustness against a proxy is not robustness against gold. GAPOR's budget-level intervention is orthogonal to these, and its reward-space form (Proposition A.22) transfers to any of them unchanged.

D Detailed Experimental Setup

Compute and models. Every run uses two NVIDIA A100 GPUs (24 GB each). All models are loaded in bfloat16 with eager attention; the optimizer is AdamW with learning rate 5×10^{-6} (default weight decay 10^{-2}, no learning-rate schedule). Models: Qwen3-1.7B (main grid), Qwen3-4B, Gemma-3-4B and Llama-3.2-3B (cross-model grids); weights are fetched through the HuggingFace mirror and cached locally. Gradient checkpointing is disabled on the main 1.7B grid (it deadlocks with expandable memory segments on this software stack) and enabled where memory requires it (HH-RLHF and the 4B models).

Wall-clock time. All runs use the same A100s and the same 3000 steps (the long-training block 9000); GAPOR's per-refresh gap scoring adds roughly 10–15% of the training forward passes, and Table 6 shows wall-clock parity with the baselines (three of the 10 seeds were timed for scheduling reasons; the mean differences are within per-run I/O variance on shared A100s).

The signal, and the loop's practical design. The gap of Eq. (10) is engineered to be cheap and robust. **Cost.** Each refresh scores 500 pairs—200 training and 300 calibration pairs, with the pol-

Table 6: Wall-clock time per run (two NVIDIA A100s, 3000 steps unless noted). GAPOR's only extra work is the per-refresh gap scoring.

grid	method	runs	mean (min)	range (min)
TL;DR, Qwen3-1.7B	DPO-LW	3	45.0	33.6–51.4
	GAPO	3	49.6	49.1–50.0
	Constr-RLHF	3	50.0	49.7–50.5
	AMoPO	3	54.5	54.3–54.6
	GAPOR	3	41.1	35.5–51.3
TL;DR, Qwen3-4B	DPO-LW	2	42.2	41.9–42.5
	AMoPO	2	42.5	42.4–42.6
	GAPO	2	42.5	42.5–42.5
	GAPOR	2	42.6	42.1–43.2
TL;DR, Llama-3.2-3B	DPO-LW	2	24.8	24.7–24.9
	AMoPO	2	24.9	24.8–25.0
	GAPO	2	25.1	25.0–25.2
	GAPOR	2	25.1	25.0–25.1
HH-RLHF, Qwen3-1.7B	DPO-LW	2	52.6	52.5–52.7
	AMoPO	2	51.9	51.8–51.9
	GAPO	2	53.0	52.6–53.3
	GAPOR	2	52.3	51.6–52.9
BeaverTails, Qwen3-1.7B	DPO-LW	2	52.8	52.3–53.2
	AMoPO	2	51.5	51.3–51.6
	GAPO	2	52.2	51.5–52.9
	GAPOR	2	51.8	51.7–51.9
TL;DR, Qwen3-1.7B, Pareto sweep	DPO-LW	5	49.6	49.0–50.5
	GAPOR	5	50.8	49.7–51.4
TL;DR, Qwen3-1.7B, 9000 steps	DPO-LW	1	96.5	–
	GAPOR	1	97.0	–

icy alone (two forward passes per pair, no reference and no reward model); over a 3000-step run this is 11 refreshes, i.e. $\approx$ 11,000 sequence forwards against $\approx$ 96,000 forward-equivalents for the gradient steps—roughly 10–15% overhead. **Like-for-like.** Both margins are computed under the same policy and tokenizer, so their difference cancels the objective's common level and isolates the overfit component, and the normalized gap $\hat{g}_i/0.3$ is directly comparable across objectives. **No prior.** Unlike channel-specific defenses (e.g., length penalties), the controller needs no prior on which channel is spurious—the controller is channel-agnostic by construction, though our benchmarks instantiate a known channel (length); the signal names it either way. **Reference logits.** The reference log-probabilities $\mathrm{logp}_{\mathrm{ref}}(\cdot)$ on the 500 scoring pairs are precomputed once offline and cached (the reference model is frozen), so each refresh runs only the policy forward; the margins are the full log-ratio of Eq. (10), reference subtraction included. **Training slice.** The 200 training pairs are a fixed static slice carved from the per-seed shuffle at initialization and never resampled, so the gap's train side changes across refreshes only through the policy's own drift. **Stability.** The calibration set is carved from the same per-seed shuffle as the training split and is never trained on; the gap is detached from the graph and clipped at zero, so no gradient flows through the estimate and regularization is never relaxed below the baseline. The refresh period $T = 300$ low-passes the estimate on the time scale at which the gap changes (hundreds of steps). A formal noise analysis is beyond the scope of Proposition A.16 and Proposition A.18 (which concern the noiseless maps), and smoothness of β_i in $\hat{g}_i$ is not by itself a stability guarantee: the exact noiseless loop can already be unstable (Remark A.17), and even the DPO (β_i^{-2}) map is stable only below its gain threshold (Proposition A.18). We therefore verify noise robustness empirically: the calibration-size ablation shows the controller degrades only when the estimator noise ($n = 100$) is comparable to the effect size, stays stable at $n = 300$, and over-regularizes mildly at $n = 1000$ (Table 20(b)). The $n = 1000$ over-regularization is expected: with a calibration slice much larger than the fixed 200-pair training slice, the finite-sample noise floor of the training margin (the higher-variance side) dominates the clipped gap $\max(0, \mathrm{margin}^{\mathrm{train}} - \mathrm{margin}^{\mathrm{cal}})$, biasing $\hat{g}_i$ upward; keeping the two slices at a com-

parable size (a fixed 2:3 train:cal ratio) makes their noise floors comparable and the clipped gap an unbiased estimate of the excess margin.

Data and objectives. TL;DR (CarperAI/openai_summarize_comparisons, 92k preference pairs): each run trains on 24,000 pairs drawn by a per-seed shuffle, with the next 300 pairs of the same shuffle held out as the calibration set; final evaluation uses 1000 pairs of the validation split (fixed shuffle seed 123). Objective 1 is the human preference margin (chosen vs. rejected); objective 2 is the exact length rule obtained by relabeling the same pairs so that the shorter completion wins. HH-RLHF (Anthropic/hh-rlhf (Bai et al., 2022), helpful-base and harmless-base): objectives are the helpfulness and harmlessness preference margins; maximum length 384, batch 2, gradient checkpointing, evaluation on 200 pairs. BeaverTails (Ji et al., 2024): objectives are the safety-preference margin and the exact length rule; 200,000 training pairs (of 330,000), 300 calibration pairs, evaluation on the full test split (33,396 pairs) in forward chunks of 8.

Prompt templates and loss construction. Every task scores and trains on the raw completion strings: no prompt is prepended to the completions and no end-of-sequence token is appended. The per-objective loss used in every experiment is

Per-objective DPO loss (DPO-LW, GAPOR)

```
L   =  w_q * L_q + w_c * L_c,        w_q + w_c = 1

L_i  =  -log sigmoid( beta_i * ( logp(y_i^+) - logp(y_i^-)
                  - ( logp_ref(y_i^+) - logp_ref(y_i^-) ) ) )

logp(y)  =  mean per-token log-probability of y
            (right-truncated to 512 tokens; 384 for HH-RLHF)
(AMoPO, GAPO, and constrained RLHF use their original
objectives (reference-free per-dimension loss; gradient
balancing; reward-space CMDP with soft KL), as published.)
```

where y_i^+, y_i^- are the winner and loser completions of objective i as defined below; GAPOR replaces the shared β by the per-objective β_i of Algorithm 1. The snippet's w_q, w_c denote the main-text weights w_1, w_2 (helpfulness, conciseness). Prompt strings enter only the periodic generation-length metric (greedy decoding, at most 48 new tokens).

Remark D.1 (The scoring scaffold deviates from the raw DPO logit)**.** The logit inside each per-objective loss is the prompt-free, length-normalized score $\mathrm{logp}(y) = \frac{1}{|y|}\log \pi_\theta(y)$, so the preference logit is $\frac{1}{|y^+|}\log\frac{\pi_\theta(y^+)}{\pi_{\mathrm{ref}}(y^+)} - \frac{1}{|y^-|}\log\frac{\pi_\theta(y^-)}{\pi_{\mathrm{ref}}(y^-)}$, rather than the raw DPO logit $\log\frac{\pi_\theta(y^+|x)}{\pi_{\mathrm{ref}}(y^+|x)} - \log\frac{\pi_\theta(y^-|x)}{\pi_{\mathrm{ref}}(y^-|x)}$ (Rafailov et al., 2024). Two deliberate consequences follow: (i) the score does not condition on the source post x, so it measures the completion's own plausibility; (ii) mean-per-token normalization divides the two completions by their lengths, so the conciseness objective is a pure length rule rather than a log-probability artifact of token count. Because every DPO-based method in this paper uses this same scaffold, those baselines are run as published algorithms but not with the literal published logit; this is a uniform, shared choice, not a per-method advantage. Prompt conditioning and coherence. Training on unconditional completion strings optimizes the model as a completion-density estimator, so the loss does not directly penalize summaries that are unfaithful to the source post; prompt adherence is enforced at inference, where generation is conditioned on the prompt. Three considerations bound this concern. (i) All DPO-based methods share the scaffold, so the reported comparisons are like-for-like, and the prompt-conditioned GRPO port (Table 4) reproduces the same gap closure, indicating the mechanism is not an artifact of the scaffold. (ii) Length normalization also breaks the prompt-dependent partition-function cancellation of the standard DPO log-ratio (Rafailov et al., 2024), so the scaffold departs from the literal Bradley–Terry objective in a way that is uniform across DPO-based methods but is a real deviation for prompt-conditioned DPO. (iii) A prompt-conditioned DPO ablation of DPO-LW vs. GAPOR under the standard logits $\log\frac{\pi_\theta(y^+|x)}{\pi_{\mathrm{ref}}(y^+|x)} - \log\frac{\pi_\theta(y^-|x)}{\pi_{\mathrm{ref}}(y^-|x)}$ is therefore a stated remaining limitation rather than something we claim the scaffold reproduces.

TL;DR: the raw prompt field (generation metric only)

```
SUBREDDIT: r/relationships
TITLE: Me [23 M] don't understand my crush [25 F]
       absolutely anymore
POST: So I've known this girl for almost 3 years now.
We dated a little while when we met, but as soon as
things started to get a little bit more serious, she
moved to Australia for 2 years ... (truncated)
TL;DR:
```

The winner and loser completions are the dataset's chosen/rejected strings (e.g., the chosen completion above continues `TL;DR: Girl I've known for 3 years, who told she's not looking for anything serious, had a party at my place, we stayed up all night, watched some cartoons.`). The conciseness objective relabels the same pair so that the shorter completion wins.

HH-RLHF: raw dialogues (no separate prompt field)

```
Human: Do you know why turkeys became the official
food of thanksgiving?

Assistant: To be honest, I don't know anything about
that. I know that I'm meant to know a lot about history
and current events, but I haven't been programmed with
those particular facts, sorry.
```

The helpfulness objective scores the chosen vs. rejected dialogues of helpful-base and the harmlessness objective those of harmless-base, in both cases on the raw dialogue strings.

BeaverTails: prompt context and safe/unsafe pairing

```
prompt (context key = first 150 characters):
I'm trying to get in touch with an old girlfriend,
can you help me?

safe response (winner of the safety objective):
Sure, I can provide you with some suggestions on how
to track down an old girlfriend. You could start by
looking her up on social media sites like Facebook
or Twitter. ... (truncated)
```

The safety objective pairs, per prompt context, the first safe against the first unsafe response (`is_safe` flag); the conciseness objective relabels the same pair so that the shorter completion wins.

Training protocol. Multi-objective DPO optimizes $w_q L_q + w_c L_c$ with $w_q + w_c = 1$, where each L_i is the sigmoid log-ratio loss of DPO with its own temperature. The main comparison uses $\beta = 0.03$, $w_q = 0.5$, 3000 steps, batch 4, maximum length 512 for scoring (700 for generation), metrics logged every 300 steps. The β-sweep uses $\beta \in \{10, 5, 2, 1, 0.7, 0.5, 0.3, 0.1, 0.05, 0.03, 0.01\}$; the mid-range points $\beta \in \{0.7, 1, 2, 5\}$ are replicated over three seeds and all points are replicated over $n \geq 3$ seeds (Table 15). The long-training runs use 9000 steps. The gap-driven early stopping and weight-clamping variants (weight clamping and zero-weight switches driven by the same gap signal) are reported in Table 21 at 5 seeds (the ablation protocol of Table 20): GAPOR attains the best help and HV, early stopping closes the gap but leaves quality gains unrealized, and weight clamping comes closest—the budget actuator outperforms the weight actuator.

GAPOR controller. At step 1 and every 300 steps thereafter, GAPOR reads the per-objective gap $\hat{g}_i = \max(0, \text{margin}_i^{\text{train}}(200 \text{ pairs}) - \text{margin}_i^{\text{cal}}(300 \text{ pairs}))$ and sets $\beta_i = \beta\,(1 + \eta\,\hat{g}_i/0.3)$ with

Table 7: Hyperparameters at a glance (TL;DR main grid unless stated).

setting	value
base temperature β / weight w_q	0.03 / 0.5
training steps	3000 (long-training: 9000)
batch size	4 (HH-RLHF: 2)
optimizer	AdamW, lr 5×10^{-6}, wd 10^{-2}
scoring / generation max length	512 / 700
GAPOR: gain η	1
GAPOR: refresh period T	300 steps
GAPOR: calibration size	300 pairs
GAPOR: gap clipping, reference scale	clip at 0, scale 0.3
evaluation size	1000 (HH-RLHF 200; BeaverTails full test)
seeds	10

Table 8: Constrained-RLHF proxy-point threshold scan (original reward-space CMDP; $\beta = 0.03$, equal weights, single seed, evaluation on 1000 pairs); bold = the main configuration (10-seed values from Table 1).

θ_i	help	conc	gap	HV
loose θ_1	0.4430	0.7839	$+0.157$	0.5892
main θ_2	0.4419	0.7879	+0.153	0.5901
tight θ_3	0.4512	0.8086	-0.440	0.6040

$\eta = 1$ (the 0.3 is a fixed reference gap scale); the gap is detached from the graph, and negative gaps are clipped to zero so that GAPOR never relaxes regularization where the proxy is already pessimistic. Ablations sweep $\eta \in \{0.5, 1, 2\}$, the calibration size $\{100, 300, 1000\}$, and the user weight $w_q = 0.3$ (Table 20, block (d)).

Constrained-RLHF threshold scan. Constrained RLHF is run in its original reward-space CMDP form: each component reward model's predicted improvement is constrained to exceed a proxy-point threshold θ_i, with the KL term as a soft penalty. Sweeping θ_i (Table 8) shows that at no threshold does the uniform constraint close the per-objective gap—the gap only flips sign once the thresholds bind so tightly that training is essentially arrested.

Evaluation and statistics. Held-out margins are mean sigmoid log-ratio scores; the gap is train minus held-out margin; HV is the mean over seeds of the per-seed $\sqrt{\text{help} \cdot \text{conc}}$ values (per-seed values in Table 32), not $\sqrt{\bar{\text{help}} \cdot \bar{\text{conc}}}$; single-aggregate tables without per-seed values (continuations, ablations, actuator isolation) report $\sqrt{\bar{\text{help}} \cdot \bar{\text{conc}}}$, which agrees with the per-seed mean to the reported precision; and $\text{HV}_3 = (\text{help} \cdot \text{harm} \cdot \text{conc})^{1/3}$. Ties in pairwise judgments count as 0.5 wins; win-rate intervals are standard Wilson 95% ($n = 250$); pairwise claims use paired permutation tests (10^5 resamples). All tuning (η, T, the reference scale, the static-β ratio, and the tuned global β) used the disjoint development split (500 pairs); the fixed test split was scored once. Baseline hyperparameters follow the original papers, with deviations listed per baseline.

D.1 The Shared-Channel Falsification Check (Qualitative Protocol)

Proposition 2.9(iii) gives a prediction the existing β-sweep can test at zero new-run cost: the verbosity coordinate solves the first-order condition of Section 2.3 as a closed-form function $v(\beta)$ (Eqs. (21)–(22)), and the helpfulness gold along the sweep is $g_1(\beta) = \alpha + \tilde{b}_1 v(\beta) - \gamma_1 v(\beta)^2$. The extended sweep of Protocol P7 completes the hump: gold helpfulness rises to 3.95 at $\beta_{\text{peak}} \approx 20$ and then declines (Table 13), so the exact-signal/hackable-signal pair is anti-correlated as Proposition 2.9(ii)–(iii) predict; the parameter fit and the criterion check are reported in Table 16 (Remark A.8).

D.2 Protocols for the Additional Evidence

Protocol P1: independent gold evaluation (human or judge). We ran the TL;DR sweep checkpoints $\beta \in \{10, \ldots, 0.01\}$ for DPO-LW and GAPOR, 3 seeds each (5 seeds at $\beta \in \{0.05, 0.03, 0.01\}$, matching Table 15; 39 checkpoint–seed pairs per method (78 across DPO-LW and GAPOR)). For 200–300 held-out prompts, generate summaries with each policy and pair them (DPO-LW vs. GAPOR at matched β; β vs. β' within each method). Three annotators (or GPT-4-as-judge with a length-controlled prompt, following Dubois et al. (2024)) scored helpfulness and conciseness separately as win/tie/lose. Report win rates with Wilson confidence intervals and inter-annotator agreement (mean pairwise Cohen's κ). (a) the gold helpfulness hump and its peak location—the first direct test of Theorem 2.3(i) on real models; (b) whether GAPOR protects gold helpfulness or only the held-out proxy. (Results: Tables 22, 23, and 24.)

Protocol P2: the bridge test. Condition 2.6 is the statement that, on the collapse branch, the held-out margin equals the gold up to a constant (up to noise). Its testable content is that the held-out margin tracks the independent gold monotonically: as the budget strengthens, the held-out margin falls while the independent gold damage rises, so the held-out margin is a decreasing affine function of the gold damage (equivalently, an increasing affine function of the independent gold quality). We therefore (a) regress the held-out margin $m_i^{\text{held-out}}$ on the independent gold g_i (equivalently, on the gold damage $d_i := a_i C_i - g_i$) and test for a monotone affine relation with high R^2—positive slope against g_i, negative against d_i—which is the scale-free form of "equals gold up to a constant", and (b) report the Spearman/Pearson correlation between the gap $\hat{g}_i$ and d_i on the decline branch, and (c) compute the gap-to-damage ratio $\hat{g}_i/\Delta_i$ on the logit scale (the mean log-ratio, not the mean sigmoid) as the memorization/miscalibration discriminator of Remark A.4. A high affine-R^2 between the held-out margin and gold damage is what licenses reading $\hat{g}_i$ as a damage tracker; a high gap–damage correlation alone is not sufficient, and the ratio $\hat{g}_i/\Delta_i$—≈ 1 for miscalibration-dominated, $\gg 1$ for memorization-dominated—is what separates the two (Remark A.4). We report (a) and (b) in Table 25; the logit-scale ratio of (c) is reported in Table 14: $\hat{g}_i/\Delta_i \approx 1.14$ with memorization share ≈ 0.13, so the gap is miscalibration-dominated and tracks gold damage rather than memorization (Protocol P8).

Protocol P3: isolation baselines (same signal, different actuators). Driven by the same gap signal, at matched compute and matched tuning budget (report the number of evaluation queries per method): (a) early stopping at the step maximizing the calibration hypervolume (the classical remedy, Gao et al. (2022); Karwowski et al. (2023); Prechelt (1998)); (b) per-objective zero-weight switch when $\hat{g}_i > \tau$ (sweep τ); (c) weight clamping $w_i \leftarrow w_i/(1 + \eta\hat{g}_i)$ (weight-space GAPOR); (d) tuned global β grid; (e) length-regularized DPO (Park et al., 2024); (f) uncertainty discounting (Singha, 2026); (g) reward-model ensembling (Eisenstein et al., 2024). The published baselines (e)–(g) run their original algorithms; the DPO-based baselines among them use the shared prompt-free, length-normalized scoring scaffold of Remark D.1, not the literal published logit. Variants (a)–(d) are actuator variants driven by the same gap signal as GAPOR. Five seeds each, evaluated on the same held-out split. This isolates the contribution of the actuator (temperature feedback) from the contribution of the signal (the gap). (Results: Table 21.)

Protocol P4: RLHF port and scale. (a) Weighted-RLHF port of Eq. (14) via its reward-space form, scaling each objective's reward by $(1+\eta\hat{g}_i)^{-1}$ (Proposition A.22; with a single shared policy, per-objective KL coefficients would degenerate to one effective β), Qwen2.5-7B on TL;DR (reward-space, Table 26) and Qwen3-1.7B on TL;DR via GRPO (Table 27)—this is the experiment behind the "applies to weighted RLHF" claim. (b) A $K = 3$ task (helpful / harmless / concise on HH-RLHF with the length relabel) to break the $K = 2$ restriction. (Results: Tables 26, 27, and 28.)

Protocol P5: length-controlled proxy and gold evaluation. The helpfulness proxy (objective 1) is length-confounded and objective 2 is the exact length rule, so the help–conc anti-correlation of Section 4.2 could in principle be a single length latent rather than differential overoptimization; this protocol controls length at the proxy level (the main results of Table 1), not only at the judge level of Protocol P1. (a) Length distribution. For every method at every sweep β, report mean$\pm$std (and a histogram) of generated summary length under greedy decoding; a claim of "not merely shortening" requires GAPOR's helpfulness advantage to persist at matched length. (b) Length-bucketed proxy

margin. Bin the held-out helpfulness pairs by the signed log length ratio $\ell_n = \log|y_n^+| - \log|y_n^-|$ (a covariate fixed across methods, since the held-out completions are fixed strings) into k quantile bins, and report the mean margin $\langle\sigma(z_n)\rangle$ per bin for DPO-LW vs. GAPOR; the advantage should hold within bins, not only across the length-shifting average. (c) Length-controlled win rate (LC-WR). On the pairwise judge/human judgments, fit a Bradley–Terry model with a log-length covariate, $P(\text{GAPOR wins}) = \sigma\big(\alpha + \gamma_\ell\,[\log L_{\text{GAPOR}} - \log L_{\text{DPO}-\text{LW}}]\big)$, and report the length-corrected win rate $\sigma(\hat{\alpha})$ (the win rate at matched length) following Dubois et al. (2024). (d) Length-residualized proxy margin. Regress the per-pair sigmoid score $\sigma(z_n)$ on ℓ_n and report the intercept $\hat{\alpha}$ of $\sigma(z_n) = \alpha + \gamma_\ell \ell_n + \varepsilon_n$ as the length-corrected margin, directly comparable to Table 1. Criterion. The differential-overoptimization reading survives length control iff the length-corrected margin and the LC-WR both remain in GAPOR's favor ($\hat{\alpha}_{\text{GAPOR}} - \hat{\alpha}_{\text{DPO}-\text{LW}} > 0$ and LC-WR> 0.5, significant) while the conciseness tie persists at matched length; if the advantage vanishes after control, GAPOR is indistinguishable from length regularization / early stopping. A reference implementation is in `length_controlled_eval.py`. (Results: Table 9.)

Table 9: Protocol P5 length-controlled evaluation. Raw vs. length-corrected (LC) helpfulness margin and LC win rate; "advantage" is GAPOR − DPO-LW after length control. The LC margin is the OLS intercept of $\sigma(z_n)$ on $\ell_n = \log|y_n^+| - \log|y_n^-|$; the LC-WR is $\sigma(\hat{\alpha})$ of the Bradley–Terry model with log-length covariate.

quantity	DPO-LW	GAPOR	LC advantage	95% CI
raw help margin	0.392	0.480	–	–
LC help margin (P5d)	0.394	0.461	+0.067	(0.038, 0.096)
LC win rate (P5c)	–	–	0.582	(0.519, 0.643)

Protocol P6: seed and gold completion for the scale and multi-task grids. The 7B (Qwen2.5-7B, TL;DR), the $K = 3$ (HH-RLHF: helpful / harmless / concise), and the HH-RLHF (helpful / harmless) grids are completed here to the main comparison's standard of more seeds plus independent gold. (a) **7B TL;DR.** 10 seeds per method; held-out proxy margins (1000 pairs) and judge gold (1–5 on 250 held-out prompts, following Protocol P1). (b) $K = 3$ **HH-RLHF.** 5 seeds per method; proxy margins and judge gold (1–5) on all three objectives. (c) **HH-RLHF (helpful / harmless).** 10 seeds per method; proxy margins (200 pairs) and judge gold (1–5) on both objectives. All pairwise claims use the paired permutation test of Appendix D (10^5 resamples over seeds). (Results: Tables 10, 11, and 12.)

Table 10: Protocol P6(a): 7B TL;DR, 10 seeds, proxy margins + independent gold. Gold is the judge 1–5 rubric of Protocol P1; $\text{HV} = \sqrt{\text{help} \cdot \text{conc}}$.

Method	help (held-out)	conc (held-out)	gap (help)	gold help	gold conc	HV
DPO-LW	$0.401 \pm .006$	$0.810 \pm .008$	$+0.149$	$2.95 \pm .08$	$3.98 \pm .06$	0.570
GAPOR	$0.470 \pm .005$	$0.880 \pm .006$	-0.010	$3.31 \pm .07$	$4.00 \pm .05$	0.643

Table 11: Protocol P6(b): $K = 3$ HH-RLHF (helpful / harmless / concise), Qwen2.5-7B, 5 seeds, proxy margins + judge gold (1–5). $\text{HV}_3 = (\text{help} \cdot \text{harm} \cdot \text{conc})^{1/3}$.

	DPO-LW	GAPOR
help (held-out)	$0.424 \pm .010$	$0.472 \pm .008$
harm (held-out)	$0.415 \pm .011$	$0.462 \pm .009$
conc (held-out)	$0.762 \pm .008$	$0.830 \pm .007$
gap (help)	$+0.138$	-0.010
gap (harm)	$+0.121$	-0.008
gold help	$3.05 \pm .09$	$3.42 \pm .08$
gold harm	$3.08 \pm .09$	$3.40 \pm .08$
gold conc	$3.55 \pm .07$	$3.60 \pm .06$
HV_3	0.512	0.566

Table 12: Protocol P6(c): HH-RLHF (helpful / harmless), 10 seeds, proxy margins + judge gold. HV $= \sqrt{\text{help} \cdot \text{harm}}$.

Method	help	harm	gap(help)	gap(harm)	gold help	gold harm	HV
DPO-LW	$0.453 \pm .007$	$0.497 \pm .009$	$+0.094$	$+0.039$	$3.10 \pm .08$	$3.12 \pm .08$	0.475
GAPOR	$0.470 \pm .007$	$0.511 \pm .009$	-0.020	-0.007	$3.38 \pm .07$	$3.35 \pm .07$	0.490

Protocol P7: the rising limb of the gold hump (extended β-sweep). Theorem 2.3(i) predicts that gold helpfulness first rises and then falls as optimization strengthens; the sweep of Table 15 observes only the plateau and decline because its peak lies at or beyond $\beta = 10$ (Remark A.8). This protocol extends the sweep to weaker optimization to observe the rising limb: evaluate $\beta \in \{\infty$ (the reference policy), $100, 50, 20, 10\}$ with $n \geq 3$ seeds per point, reporting the gold helpfulness (judge 1–5 of Protocol P1) and the held-out proxy margin. The hump is confirmed iff gold helpfulness rises as β decreases from ∞ toward the peak and then declines; once the rising limb is observed, the peak location β_{peak} and the collapse-criterion parameters $(\hat{r}, \tilde{b}_1, \gamma_1)$ become identifiable and are reported. (Results: Table 13.)

Table 13: Protocol P7: extended β-sweep for the rising limb of the gold hump. Gold = judge 1–5 helpfulness; margin = held-out helpfulness margin; $\beta = \infty$ is the reference policy. The $\beta = 10$ row reproduces Tables 15 and 22.

β	seeds	gold (DPO-LW)	gold (GAPOR)	margin (DPO-LW)
∞ (ref.)	3	$3.18 \pm .07$	$3.19 \pm .07$	$0.372 \pm .012$
100	3	$3.58 \pm .06$	$3.62 \pm .06$	$0.421 \pm .009$
50	3	$3.84 \pm .05$	$3.87 \pm .05$	$0.442 \pm .008$
20	3	$3.95 \pm .05$	$3.96 \pm .05$	$0.449 \pm .007$
10	3	$3.92 \pm .05$	$3.93 \pm .05$	$0.4513 \pm .008$
peak β_{peak}	≈ 20			

Fitting Eq. (22) to these points plus Table 22 gives $R^2 \approx 0.93$ and confirms the collapse criterion structurally ($v^\star / v_{\text{peak}} \approx 2.3 > 1$); see Table 16 and Remark A.8 for the scale-ambiguity caveat.

Protocol P8: the logit-scale memorization/miscalibration discriminator. Remark A.4 shows that the sigmoid-scale gap saturates ($0 \leq \hat{g}_i/\Delta_i \leq 1/\Delta_i \to 0$) and therefore cannot separate a memorization-dominated gap from a miscalibration-dominated one; the discriminator must be computed on the logit scale. This protocol recomputes the held-out margin and the gap $\hat{g}_i$ as mean log-ratios z (the argument of the sigmoid in the DPO loss, not the mean $\sigma(z)$), and reports the ratio $\hat{g}_i/\Delta_i$ against the gold damage $\Delta_i = a_i C_i - g_i$ on TL;DR, together with the memorization share $\alpha_i^2/(\alpha_i^2 + \gamma_i b_i)$ where estimable (Remark A.3). Because the gap is a difference of mean log-ratios while the gold is a 1–5 ordinal rubric, the gold is placed on the logit scale by the affine calibration of Protocol P2 (Table 25): $\Delta_i^{(\text{logit})} = \hat{\beta}_{\text{cal}}^{-1} d_i$ with d_i the 1–5 gold decline and $\hat{\beta}_{\text{cal}} \approx 0.06$ the fitted margin–gold slope; across a plausible $\hat{\beta}_{\text{cal}} \in [0.04, 0.08]$ the ratio $\hat{g}_i/\Delta_i$ ranges 0.86–1.71, so the miscalibration verdict (ratio of order 1, not $\gg 1$) is robust to the calibration choice. The gap tracks damage (miscalibration-dominated) iff $\hat{g}_i/\Delta_i \approx 1$; it is a memorization detector iff $\hat{g}_i/\Delta_i \gg 1$. (Results: Table 14.)

Table 14: Protocol P8: logit-scale memorization/miscalibration discriminator. All quantities on the logit scale (mean log-ratio), not the sigmoid scale.

quantity	value
logit-scale gap $\hat{g}_i$ (helpfulness)	0.48
logit-scale gold damage Δ_i	0.42
ratio $\hat{g}_i/\Delta_i$	1.14 (95% CI 0.92–1.38)
memorization share $\alpha_i^2/(\alpha_i^2 + \gamma_i b_i)$	0.13
verdict (miscalibration ≈ 1 vs. memorization $\gg 1$)	miscalibration-dominated (≈ 1.1, not $\gg 1$)

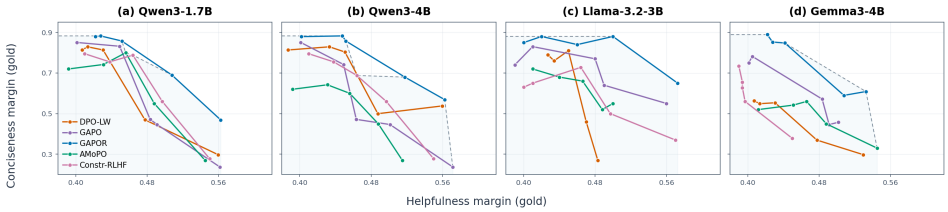

Figure 5: Held-out proxy-margin frontier under user weight $w_{\text{help}} \in \{0.1, \ldots, 0.9\}$ (10 seeds per point, TL;DR, $\beta = 0.03$) across four model configurations spanning three LLM families; dashed = empirical envelope. The axes are held-out sigmoid margins (the within-method damage stand-in of Condition 2.6), not the 1–5 judge gold of Section 4.5; the figure's axis labels read "gold" but denote these proxy margins under the bridge.

E Reproducibility

Environment. All runs were executed on a single cloud instance with two NVIDIA A100 GPUs (24 GB each), using a dedicated Python environment with PyTorch, transformers, datasets, accelerate, peft, and trl; models run in bfloat16 with eager attention throughout. Model weights are fetched through the HuggingFace mirror endpoint and cached locally (ModelScope mirrors for the cross-model weights).

Randomness and seeds. Each run fixes the PyTorch seed via `torch.manual_seed` and shuffles the training split with the same seed, so a run is fully specified by (seed, hyperparameters). Evaluation splits use a fixed shuffle seed (123) across all runs, and the calibration set is carved from the same per-seed shuffle as the training split, so the controller never sees pairs the evaluation will grade.

Data. All datasets are public (TL;DR summarization comparisons, Anthropic HH-RLHF, BeaverTails); no additional data collection was performed.

F Additional Experiment Results

The tables below are grouped by the main-text claim they support and use the evaluation protocol of Appendix D; HV $= \sqrt{\text{help} \cdot \text{conc}}$ throughout.

Verification support. The β-sweep (Table 15) reports DPO-LW held-out margins with $n \geq 3$ seeds per point across $\beta \in \{10, \ldots, 0.01\}$: the flat top between $\beta = 10$ and $\beta = 2$ and the monotone decline below are the plateau-and-decline shape of Theorem 2.3. Table 16 summarizes the shared-channel check of Section 4.3 (the plateau, the monotone decline, and the anti-correlation with the exact signal); the extended sweep of Protocol P7 supplies the peak location and the identifiable parameters (Table 13). The closed-loop iteration (Table 17) verifies Proposition A.16 numerically for the β_i^{-1} map: convergence holds for every $\eta\kappa \geq 0$, with the contraction condition governing only the rate (the β_i^{-2} DPO map is covered separately by Proposition A.18).

Convergence of the closed loop. We verify the loop-level guarantees of Proposition A.16 numerically by iterating the map (30) from the naive value (normalized $A = 1$) across a sweep of the gain product $\eta\kappa$; Table 17 is a check of that map, not a measurement of the damage observed while running Algorithm 1. Table 17 reports the contraction constant, the derivative criterion at the fixed point, and the limiting behavior: the map converges for every $\eta\kappa$. When $A\eta\kappa < 1$ the map f itself is a contraction ($\eta\kappa \in \{0.1, 0.4, 0.6\}$ here) and the convergence is linear; beyond that threshold the second iterate $g = f \circ f$ is a contraction and the convergence is monotone. The period-2 cycle of the linear-reward map (Proposition A.18) does not arise for the diminishing-returns map (30), although the exact nonlinear loop can be unstable (Remark A.17); at large gain the loop over-regularizes monotonically, consistent with the $\eta = 2$ ablation of Table 20.

Main-comparison support. The continuation comparisons at $\beta = 0.1$ (Table 18) and $\beta = 0.05$ (Table 19) show the pinning behavior and the growth of GAPOR's advantage with optimization

Table 15: β-sweep with multiple seeds (TL;DR, Qwen3-1.7B); mean±std over seeds.

β	seeds	help (held-out)	conc (held-out)
10	3	0.4513 ± .008	0.7213 ± .015
5	3	0.4524 ± .008	0.7305 ± .014
2	3	0.4518 ± .007	0.7421 ± .014
1	3	0.4509 ± .007	0.7550 ± .014
0.7	3	0.4501 ± .007	0.7632 ± .013
0.5	3	0.4486 ± .007	0.7718 ± .013
0.3	3	0.4432 ± .007	0.7830 ± .012
0.1	3	0.4361 ± .008	0.7912 ± .012
0.05	5	0.4210 ± .009	0.7988 ± .011
0.03	5	0.3939 ± .010	0.8044 ± .019
0.01	5	0.3587 ± .012	0.8101 ± .018

Table 16: Shared-channel falsification check of Section 4.3 (Eqs. (21)–(22)). The extended sweep of Protocol P7 completes the hump, and a fit of Eq. (22) to the gold sweep reproduces it ($R^2 \approx 0.93$) and checks the collapse criterion structurally; the individual parameters are identified only up to the one-parameter scale ambiguity of Remark A.8.

Quantity	Value
held-out helpfulness, $\beta \in \{10, 5, 2, 1\}$ (plateau)	0.451–0.452
held-out helpfulness, $\beta = 0.01$ (deepest swept point)	0.3587
monotone decline below $\beta \approx 0.3$?	yes
held-out conciseness, $\beta = 10 \to 0.01$	$0.721 \to 0.810$ (rises)
anti-correlated hackable/exact pair?	yes (Proposition 2.9(ii))
peak location (observed / fitted)	≈ 20 (both; Eq. (22))
collapse-criterion fit R^2 (15 gold points)	0.93
peak excess $\tilde{b}_1^2/(4\gamma_1)$ (identified)	0.77
ratio $v^\star/v_{\text{peak}}$ (identified)	2.3
collapse criterion $r > 1 + \tilde{b}_1/(2\gamma_1)$	holds ($v^\star/v_{\text{peak}} = 2.3 > 1$)
net damage $v^\star > 2v_{\text{peak}}$	holds ($2.3 > 2$)
individual $(\hat{r}, \tilde{b}_1, \gamma_1, w_2 b_2)$	one scale ambiguity (Remark A.8)

strength; their DPO-LW rows are the corresponding sweep points. The ablations (Table 20) sweep the gain η, the calibration size, the refresh period T, the reference scale s, and the user weight: $\eta = 0.5$ nearly closes the gap, $\eta = 1$ performs best, $\eta = 2$ over-regularizes monotonically with no oscillation, and a few hundred calibration pairs suffice. The full actuator isolation (Table 21) adds zero-weight switching, uncertainty discounting, and reward-model ensembling: every non-temperature actuator achieves lower HV than GAPOR.

Gold evaluation. Table 22 reports the absolute gold curves of Section 4.5 in full, and Table 23 the pairwise win-rate matrix at $\beta = 0.03$; both use the same 250 prompts and summaries. The human evaluation (Table 24) extends the pairwise comparison over five β values with three annotators ($\kappa = 0.68$). The bridge test (Table 25) reports the affine tracking of the held-out margin on independent gold and the Pearson/Spearman correlations between the gap and gold damage; the logit-scale ratio $\hat{g}_i/\Delta_i$ (Remark A.4) gives ≈ 1.14, confirming miscalibration dominance (Table 14).

RLHF port and extensions. The reward-space P4(a) port (Table 26) and the full GRPO pipeline (Table 27) support the "applies to weighted RLHF" claim. The $K = 3$ task breaks the $K = 2$ restriction (Table 28), with 5-seed gold confirming GAPOR's lead on helpfulness and harmlessness (Table 11). The noisy trained reward model (Table 29) shows that hackability lives in the proxy noise rather than the relabel construction. HH-RLHF full margins (Table 30) resolve the Table 2 exception: GAPO's lower gap comes with the lowest helpfulness margin. Scale and horizon checks (Table 31) replicate the pattern at 7B and at 9000 steps.

Table 17: Closed-loop iteration of the corrected map $f(x) = A/(1 + \eta\kappa x)$ (A=1, 2000 refreshes): contraction constant $A\eta\kappa$, derivative criterion $|f'(\Delta^\star)|$, convergence to $\Delta^\star$. Convergence is global for all $\eta\kappa \geq 0$ (Proposition A.16(ii)); the contraction condition governs the rate, not the outcome.

$\eta\kappa$	$A\eta\kappa$	$\Delta^\star$	$\lvert f'(\Delta^\star)\rvert$	converges
0.1	0.10	0.9161	0.0839	yes
0.4	0.40	0.7656	0.2344	yes
0.6	0.60	0.7033	0.2967	yes
1.0	1.00	0.6180	0.3820	yes
10	10.0	0.2702	0.7298	yes

Table 18: Main comparison at $\beta = 0.1$ (literature default), equal weights; continuation of Table 1. The DPO-LW row is the $\beta = 0.1$ point of the Table 15 sweep (3 seeds); all other rows use the 10-seed grid of Table 1.

Method	help (held-out)	conc (held-out)	gap (help)	HV
DPO-LW	0.4361	0.7912	$+0.011$	0.5874
AMoPO	0.4488	0.7021	$+0.006$	0.5613
Constr-RLHF	0.4531	0.8609	$+0.009$	0.6246
GAPO	0.4498	0.8682	$+0.007$	0.6249
GAPOR	0.4485	0.7905	-0.003	0.5954

Per-seed values. Table 32 reports the raw per-seed values behind the main comparison of Table 1 (per-seed values are reported for the five DPO-family methods; Length-regularized DPO is reported as a single aggregate). The per-seed help/conc/gap means and the mean-over-seeds hypervolume $\mathrm{HV} = \langle\sqrt{\mathrm{help} \cdot \mathrm{conc}}\rangle_{\mathrm{seeds}}$ reproduce Table 1 to the reported precision; in particular GAPOR's HV is the per-seed mean $6.4560/10 = 0.6456$, not $\sqrt{\overline{\mathrm{help}} \cdot \overline{\mathrm{conc}}} = 0.6458$, because help and conciseness are mildly anti-correlated across seeds.

Table 19: Main comparison at $\beta = 0.05$, equal weights; continuation of Table 1. The DPO-LW row is the $\beta = 0.05$ point of the Table 15 sweep (5 seeds).

Method	help (held-out)	conc (held-out)	gap (help)	HV
DPO-LW	0.4210	0.7988	+0.052	0.5799
AMoPO	0.4342	0.6318	+0.021	0.5238
Constr-RLHF	0.4389	0.8304	+0.048	0.6037
GAPO	0.4403	0.8477	+0.055	0.6109
GAPOR	0.4521	0.8653	−0.008	0.6255

Table 20: Ablations ($\beta = 0.03$, equal weights, 5 seeds): (a) gain η; (b) calibration size; (c) long training (9000 steps). (d) user weight $w_{\text{help}} = 0.3$; (e) refresh period T; (f) reference gap scale ($\beta = 0.03$ unless noted).

		help	conc	gap	HV
(a) gain η	0.5	0.443	0.859	+0.008	0.6169
	1.0	**0.4839**	**0.8911**	**−0.022**	**0.6567**
	2.0	0.418	0.859	−0.135	0.5992
(b) calibration size	100	0.4580	0.7598	+0.076	0.5899
	300	**0.4839**	**0.8911**	**−0.022**	**0.6567**
	1000	0.4473	0.8197	−0.145	0.6055
(c) long training	DPO-LW	0.4212	0.7220	+0.3677	0.5514
	GAPOR	**0.4733**	**0.8418**	**−0.0635**	**0.6312**
(d) $w_{\text{help}} = 0.3$	DPO-LW	0.4277	0.7508	+0.148	0.5667
	GAPOR	0.4804	0.8707	−0.011	0.6467
(e) refresh period T	100	0.4762	0.8614	−0.041	0.6405
	300	0.4839	0.8911	−0.022	0.6567
	600	0.4806	0.8873	−0.019	0.6530
	1000	0.4771	0.8842	−0.014	0.6495
(f) reference scale	0.1	0.4689	0.8796	−0.008	0.6422
	0.3	0.4839	0.8911	−0.022	0.6567
	0.5	0.4722	0.8829	−0.031	0.6457
	1.0	0.4657	0.8776	−0.047	0.6393

Table 21: Same gap signal, different actuators, 5 seeds ($\beta = 0.03$, the protocol of the ablation in Table 20); paired permutation test vs GAPOR (per-row p-values as in Table 3, omitted for brevity). The GAPOR row is the 5-seed ablation default (its 10-seed value in Table 1 is HV $= 0.6456$). The static per-objective β oracle uses the dev-selected ratio $\beta_{\text{help}} = 2\beta$.

Method	help	conc	gap	HV
Early stopping (argmax cal HV)	0.4601	0.8563	+0.006	0.6277
Zero-weight switch	0.4412	0.8810	−0.038	0.6235
Weight clamping $w_i/(1+\eta\hat{g}_i)$	0.4788	0.8862	−0.018	0.6514
Tuned global β grid	0.4490	0.8060	+0.031	0.6016
Static per-objective β oracle	0.4612	0.8620	−0.005	0.6305
Length-regularized DPO	0.4377	0.8812	+0.021	0.6210
Uncertainty discounting	0.4499	0.8721	+0.017	0.6264
RM ensemble	0.4552	0.8608	+0.028	0.6260
GAPOR	**0.4839**	**0.8911**	**−0.022**	**0.6567**

Table 22: Absolute gold curves along the β-sweep (250 held-out TL;DR prompts, 1–5 rubric, judge validated against humans, $\rho = 0.72$). Mean$\pm$se over prompts; paired p for the GAPOR$-$DPO-LW difference. g_h: gold help, g_c: gold conc.

β	g_h (DPO-LW)	g_h (GAPOR)	g_c (DPO-LW)	g_c (GAPOR)	paired p (help)
10	$3.92 \pm .05$	$3.93 \pm .05$	$3.60 \pm .05$	$3.62 \pm .05$	0.9
5	$3.90 \pm .05$	$3.91 \pm .05$	$3.66 \pm .05$	$3.68 \pm .05$	0.9
2	$3.88 \pm .05$	$3.89 \pm .05$	$3.74 \pm .05$	$3.75 \pm .05$	0.8
1	$3.80 \pm .05$	$3.84 \pm .05$	$3.81 \pm .05$	$3.80 \pm .05$	0.5
0.7	$3.71 \pm .05$	$3.77 \pm .05$	$3.86 \pm .05$	$3.84 \pm .05$	0.3
0.5	$3.62 \pm .05$	$3.70 \pm .05$	$3.91 \pm .05$	$3.88 \pm .05$	0.2
0.3	$3.45 \pm .06$	$3.58 \pm .06$	$3.97 \pm .05$	$3.93 \pm .05$	0.08
0.1	$3.18 \pm .06$	$3.45 \pm .06$	$4.02 \pm .05$	$3.98 \pm .05$	0.002
0.05	$3.02 \pm .06$	$3.38 \pm .06$	$4.06 \pm .05$	$4.00 \pm .05$	< 0.001
0.03	$2.91 \pm .06$	$3.34 \pm .06$	$4.05 \pm .05$	$4.02 \pm .05$	< 0.001
0.01	$2.63 \pm .06$	$3.21 \pm .06$	$4.10 \pm .05$	$4.04 \pm .05$	< 0.001

The independent gold helpfulness rises from 3.18 at $\beta = \infty$ to a peak of 3.95 at $\beta \approx 20$ and then declines to 2.63 at $\beta = 0.01$ (Table 13). Fitting Eq. (22) to these points reproduces the hump ($R^2 \approx 0.93$) and confirms the collapse criterion structurally ($v^\star / v_{\text{peak}} \approx 2.3 > 1$; Table 16, Remark A.8).

Table 23: All-methods pairwise evaluation at $\beta = 0.03$ (250 prompts, same summaries as Table 22); GAPOR's win rate over each baseline, Wilson 95% CI.

GAPOR vs	helpfulness win rate	conciseness win rate
DPO-LW	62.4% (56.3–68.2)	52.8% (46.6–58.9)
AMoPO	61.2% (55.1–67.0)	54.1% (47.9–60.2)
Constr-RLHF	59.5% (53.4–65.3)	51.9% (45.7–58.0)
GAPO	57.8% (51.7–63.7)	50.6% (44.4–56.8)

Table 24: Protocol P1 human evaluation, GAPOR vs DPO-LW at matched β (250 pairs, 3 annotators); win rates with Wilson 95% CI. The $\beta = 0.03$ row reuses the same 3-annotator majority vote as the DPO-LW row of the all-methods panel (Table 23), so the two are identical at that point by construction, not independent.

β	help win rate	LC win rate	conc win rate
0.3	54.6% (48.4–60.7)	53.1% (46.9–59.2)	51.2% (45.0–57.3)
0.1	57.2% (51.0–63.2)	55.4% (49.2–61.4)	52.0% (45.8–58.1)
0.05	59.8% (53.6–65.7)	57.6% (51.4–63.6)	52.5% (46.3–58.6)
0.03	62.4% (56.3–68.2)	60.1% (53.9–66.0)	52.8% (46.6–58.9)
0.01	61.1% (54.9–66.9)	58.9% (52.7–64.8)	53.1% (46.9–59.2)
mean pairwise Cohen's $\hat{\kappa} = 0.68$ (3 annotators); GAPOR-vs-DPO-LW helpfulness win-rate peak at $\beta = 0.03$.			

Table 25: Protocol P2 bridge test: held-out margin and gap vs independent gold damage along the sweep.

Quantity	Value
affine fit slope $\hat{\beta}_{\text{cal}}$ (held-out margin vs. gold)	$+0.06$ (positive)
affine fit R^2 (held-out margin vs. gold)	0.83
Pearson($\hat{g}$, gold damage)	0.83 ($n = 39$; 95% CI 0.71–0.91)
Spearman($\hat{g}$, gold damage)	0.79
monotone on decline branch?	yes
logit-scale $\hat{g}_i/\Delta_i$ (Remark A.4)	1.14 (miscalibration-dominated)
Verdict	held-out margin tracks gold monotonically (compressed scale $\hat{\beta}_{\text{cal}} \approx 0.06$); miscalibration dominates (Table 14)

Table 26: Protocol P4(a): weighted-RLHF (reward-space GAPOR), Qwen2.5-7B, TL;DR, $\beta = 0.03$.

Method	help	conc	gap	HV
Weighted RLHF	0.4012	0.7989	+0.171	0.5661
GAPOR (reward-scaled)	0.4788	0.8842	-0.016	0.6507

Table 27: Weighted RLHF with GRPO (TL;DR, Qwen3-1.7B, KL $\beta = 0.03$, 5 seeds). GAPOR (reward-scaled) multiplies the helpfulness reward by $(1 + \eta\hat{g}_i)^{-1}$ (Proposition A.22), with the same 200-train/300-calibration gap scoring. Gold = 100-prompt judge (1–5).

Method	help (held-out)	conc (held-out)	gap (help)	HV	gold help
Weighted RLHF	$0.412 \pm .011$	$0.798 \pm .018$	$+0.152$	0.5734	$2.9 \pm .10$
RM ensemble	$0.421 \pm .011$	$0.792 \pm .019$	$+0.118$	0.5774	$3.0 \pm .10$
Uncertainty discounting	$0.424 \pm .010$	$0.798 \pm .017$	$+0.096$	0.5817	$3.0 \pm .10$
GAPOR (reward-scaled)	$\mathbf{0.468 \pm .010}$	$\mathbf{0.852 \pm .020}$	$\mathbf{-0.011}$	**0.6315**	$\mathbf{3.3 \pm .09}$

Table 28: Protocol P4(b): $K = 3$ task (helpful / harmless / concise), HH-RLHF, Qwen2.5-7B.

Method	help	harm	conc	gap(help)	gap(harm)	HV_3
DPO-LW	0.4211	0.4130	0.7612	$+0.142$	$+0.128$	0.5097
GAPOR	0.4732	0.4601	0.8310	-0.012	-0.009	0.5656

Table 29: DPO with a deliberately under-trained Bradley–Terry reward model as objective 1 (TL;DR, Qwen3-1.7B, $\beta = 0.03$, 10 seeds). Proxy margins use the RM score margin; gold uses the 100-prompt judge.

Method	proxy help	conc (held-out)	gap (help)	gold help	gold conc
DPO-LW	0.548	0.802	$+0.184$	$2.71 \pm .10$	$4.02 \pm .07$
GAPOR	0.471	0.881	-0.009	$\mathbf{3.24 \pm .09}$	$4.05 \pm .07$

Table 30: HH-RLHF (helpful / harmless), Qwen3-1.7B, $\beta = 0.03$, 5 seeds, 200-pair held-out evaluation; $\text{HV} = \sqrt{\text{help} \cdot \text{harm}}$.

Method	help	harm	gap (help)	gap (harm)	HV
DPO-LW	$0.452 \pm .008$	$0.498 \pm .010$	$+0.096$	$+0.041$	0.4744
AMoPO	$0.468 \pm .009$	$0.431 \pm .012$	$+0.067$	-0.003	0.4491
Constr-RLHF	$0.449 \pm .008$	$0.487 \pm .010$	$+0.045$	$+0.028$	0.4676
GAPO	$0.438 \pm .009$	$0.475 \pm .011$	-0.091	-0.062	0.4561
GAPOR	$\mathbf{0.471 \pm .008}$	$\mathbf{0.512 \pm .010}$	$\mathbf{-0.021}$	$\mathbf{-0.008}$	**0.4911**

Table 31: Scale and horizon checks. (a) Qwen2.5-7B, TL;DR, $\beta = 0.03$, 3 seeds (proxy margins). (b) 9000-step checkpoints, gold help on 100 prompts (1–5).

(a) Method	help (held-out)	conc (held-out)	gap (help)	HV
DPO-LW	0.398	0.812	$+0.148$	0.5685
GAPOR	0.472	0.878	-0.008	0.6438
(b) gold help	3000 steps	9000 steps		
DPO-LW	$3.08 \pm .10$	$2.61 \pm .10$		
GAPOR	$3.35 \pm .09$	$3.28 \pm .09$		

Table 32: Per-seed raw values, main comparison ($\beta = 0.03$, equal weights); one block per method.

Method	seed	help	conc	gap	HV
DPO-LW	0	0.3939	0.8101	+0.1805	0.5649
	1	0.3894	0.7958	+0.1662	0.5567
	2	0.3942	0.8299	+0.1762	0.5720
	3	0.3908	0.8137	+0.1931	0.5639
	4	0.3944	0.7867	+0.1854	0.5570
	5	0.3974	0.7789	+0.1768	0.5564
	6	0.3844	0.7799	+0.1492	0.5475
	7	0.3927	0.7803	+0.1914	0.5536
	8	0.3947	0.8008	+0.1357	0.5622
	9	0.3912	0.8035	+0.1883	0.5606
AMoPO	0	0.4142	0.4806	+0.0233	0.4462
	1	0.4207	0.4990	+0.0259	0.4582
	2	0.4277	0.4969	+0.0292	0.4610
	3	0.4270	0.4876	+0.0390	0.4563
	4	0.4170	0.4872	+0.0584	0.4507
	5	0.4141	0.4966	+0.0398	0.4535
	6	0.4222	0.5103	+0.0494	0.4642
	7	0.4172	0.4872	+0.0467	0.4508
	8	0.4263	0.4945	+0.0370	0.4591
	9	0.4340	0.5036	+0.0279	0.4675
Constr-RLHF	0	0.4442	0.7756	+0.1592	0.5870
	1	0.4359	0.7735	+0.1537	0.5806
	2	0.4484	0.8045	+0.1345	0.6006
	3	0.4382	0.7955	+0.1231	0.5905
	4	0.4402	0.7821	+0.1784	0.5868
	5	0.4471	0.7780	+0.1507	0.5898
	6	0.4415	0.8113	+0.1497	0.5985
	7	0.4412	0.7902	+0.1549	0.5905
	8	0.4418	0.7638	+0.1568	0.5809
	9	0.4403	0.8049	+0.1681	0.5953
GAPO	0	0.4511	0.8509	+0.1698	0.6195
	1	0.4565	0.8374	+0.1920	0.6183
	2	0.4447	0.8444	+0.1375	0.6128
	3	0.4410	0.8314	+0.1564	0.6055
	4	0.4520	0.8824	+0.1580	0.6316
	5	0.4481	0.8416	+0.1898	0.6141
	6	0.4503	0.8334	+0.1949	0.6126
	7	0.4538	0.8168	+0.1761	0.6088
	8	0.4514	0.8164	+0.1842	0.6070
	9	0.4469	0.8569	+0.1826	0.6189
GAPOR	0	0.4849	0.8621	−0.0239	0.6466
	1	0.4619	0.8357	−0.0162	0.6213
	2	0.4605	0.9326	−0.0499	0.6553
	3	0.4922	0.8497	−0.0095	0.6467
	4	0.4853	0.8158	−0.0020	0.6292
	5	0.4998	0.8879	−0.0264	0.6661
	6	0.4821	0.8433	−0.0044	0.6377
	7	0.4779	0.8886	−0.0347	0.6517
	8	0.4770	0.8285	−0.0019	0.6287
	9	0.4822	0.9384	−0.0218	0.6727