SYNTACTIC SHADOWING: WHY LENGTH PENALTIES DO NOT FIX PREFERENCE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Preference optimization is supposed to teach a language model what people prefer, but a training pair exposes more than meaning. The resulting shortfall has been attributed to length bias and answered with length penalties and length normalization, as if controlling length were controlling non-semantic bias. These remedies have not fixed preference optimization: models trained with them align meaning no better than models trained without them. We ask whether the reason is a confound in the training signal itself. A longer response is rarely just a longer version of a shorter one; it carries different sentence structure. Length and syntax are confounded in natural text, so a model that appears to prefer long responses may in fact prefer the syntax that long responses tend to have. To make this measurable, we use meaning-preserving rewrites in which each step changes exactly one property, yielding a length gap measured at fixed syntax and a syntactic gap measured at fixed length, both read against the discrepancy between two rewrites produced under the same specification. Building on this, we propose a framework that decomposes the alignment gap of a preference pair into a length, a syntactic and a semantic component, together with a construction that rewrites the rejected response into token-exact counterfactuals of the chosen one. Applying this to 25 models spanning ten state-of-the-art methods, and to 16 models of a second backbone, reveals a consistent pattern. Whether trained or not, and whichever method is used, directional semantic preference does not track the length channel, and the syntactic channel beneath it stays where it was; what does improve it, in a controlled comparison on identical data, is training on surface-matched pairs, in which the rejected response is rewritten to the chosen one's length, or to its length and syntax. We call this *syntactic shadowing*: the syntactic component of surface bias sits in the shadow of length, where length control does not reach it.[1]

1 INTRODUCTION

Large language models (LLMs) such as GPT-4o, Llama 3 and Qwen2.5 write code, summarize documents and answer open-ended questions at a level that was until recently thought to require human judgment (Hurst et al., 2024; Grattafiori et al., 2024; Yang et al., 2024; Bubeck et al., 2023). Their usefulness rests on a post-training stage that adapts a pretrained model to what people actually want, and the dominant recipe for that stage is preference alignment. Reinforcement learning from human feedback (RLHF) fits a Bradley–Terry reward model to pairwise human judgments and fine-tunes the policy against it (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022; BRADLEY & TERRY, 1952). Direct preference optimization (DPO) collapses the two stages into one supervised objective on the log-probability ratios of the preferred and dispreferred response (Rafailov et al., 2023), and a family of variants has followed (Azar et al., 2024; Xu et al., 2024; Ethayarajh et al., 2024; Meng et al., 2024; Hong et al., 2024).

Despite this success, these methods do not align *meaning* as well as one would hope. On the TL;DR summarization preferences studied here (Völske et al., 2017; Stiennon et al., 2020), 25 models trained with ten state-of-the-art (SOTA) methods reach a pairwise accuracy of only 0.51–0.60 on

[1] Code is available at
anonymous.4open.science/r/semantic-vs-surface-preference-alignment-5FFD

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.

held-out pairs, against 0.518 untrained. When the rejected summary is rewritten to the chosen summary's length and sentence pattern, so that only meaning differs, even plain DPO prefers the chosen meaning in 79.0% of pairs, against 77.4% for the untrained model (Section 5.3). The most common explanation is *length bias*: annotators prefer longer responses, the training signal inherits the preference, and the policy learns to be verbose rather than better (Singhal et al., 2023; Dubois et al., 2024). A series of methods therefore removes length from the objective, each at a different place: R-DPO penalizes the length difference inside the DPO logit, $-\alpha(|y_w| - |y_l|)$ (Park et al., 2024); SimPO drops the reference model and rewards the per-token average log-probability $(\beta/|y|)\log \pi_\theta(y \mid x)$ (Meng et al., 2024); ORPO puts an odds-ratio penalty on length-normalized log-probabilities on top of the SFT loss (Hong et al., 2024); LD-DPO down-weights the tokens beyond the shared length (Liu et al., 2024); SamPO down-samples token log-probabilities so that both responses enter the logit with equally many tokens (Lu et al., 2024); LMPO controls length inside a reference-free margin objective (Li et al., 2025); tie training adds tied pairs to break the label–length correlation in the data (Moya et al., 2026); and distributionally robust variants such as Dr. DPO reweight pairs against such confounds (Wu et al., 2025). Reward models are treated likewise (ODIN, RRM; Chen et al., 2024; Liu et al., 2025a), as is RL (Dr. GRPO, GSPO; Liu et al., 2025b; Zheng et al., 2025).

Preference alignment is nevertheless not a solved problem, for two reasons. First, the mitigation is incomplete. Measured on meaning-preserving rewrites that differ only in length, R-DPO is *more* length-sensitive than plain DPO (1.661 vs. 1.551 in the units of Section 3), and penalties on the mean length bias (the I4 variants in Table 2, defined in Section 5.1) drive that mean to its minimum while pushing per-example length sensitivity to the highest values in our study (2.33 and 2.62). Second, and more importantly, lower length sensitivity does not translate into better semantic alignment. SimPO reduces length sensitivity to 1.362, among the lowest in our study, yet none of the seven length-handling methods, SimPO included, improves directional semantic preference over plain DPO on pairs whose length and sentence pattern are matched (Section 5.3). Length bias is therefore not the main factor limiting semantic alignment.

We therefore study one concrete question: *how much of preference alignment is actually affected by length bias, how much by the syntax that co-varies with length, and how much by semantics?* Its answer partially explains the three observations above: why meaning is learned poorly, why length penalties do not remove length bias completely, and why alignment does not improve even when they do.

Answering it requires separating length from something that almost always travels with it. A longer response is rarely just a longer version of a shorter one; it carries different sentence structure, more subordination, different clause boundaries. Length and syntax are confounded in natural text, so a model that appears to prefer long responses may in fact prefer the syntax that long responses tend to have. We call this *syntactic shadowing*: length bias, as it is commonly measured, can contain a syntactic component that co-varies with length, and controlling length is then taken to control non-semantic bias while the syntactic component stays in its shadow. Throughout, *semantics* is what a response says and *syntax* its sentence pattern (clause order, subordination, clause boundaries); mere rewording, which changes neither, is within-specification variability.

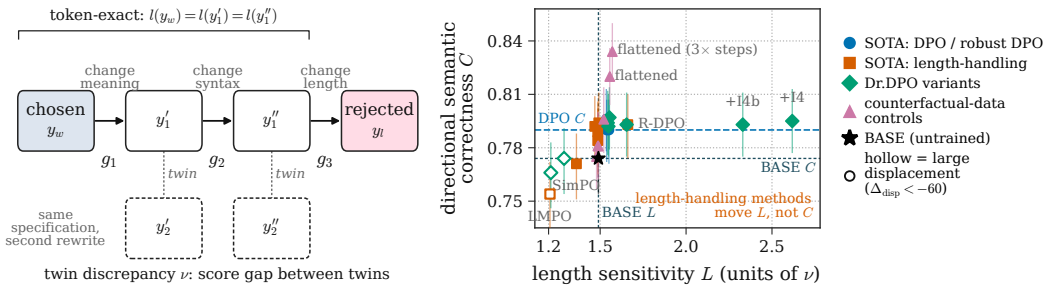
What we want preference optimization to learn is the difference in semantics between the chosen and the rejected response; what a training pair exposes is a difference in semantics, syntax and length at once. Let x be a prompt, y a response and $\log \pi_\theta(y) = \log \pi_\theta(y \mid x)$ its sequence log-probability under the model θ. For a preference pair (x, y_w, y_l) of chosen and rejected responses, the *alignment gap* is

$$G_\theta(x; y_w, y_l) \;:=\; \log \pi_\theta(y_w) - \log \pi_\theta(y_l), \tag{1}$$

the quantity that decides which response the model prefers. Pair by pair it splits exactly into a length, a residual syntactic and a residual semantic gap, $G_\theta = g_3 + g_2 + g_1$; across pairs the three gaps take both signs, so what the training signal can exploit is bounded by their magnitudes:

$$\mathbb{E}\,|G_\theta| \;\le\; \mathbb{E}\,|g_3| + \mathbb{E}\,|g_2| + \mathbb{E}\,|g_1|. \tag{2}$$

The rewrites that define the three gaps act on the rejected response alone. The chosen response is the good data, the target the pair is meant to teach, so we hold it fixed and ask of the rejected response why it is rejected: for its length, for its sentence pattern, or for what it says. Each rewrite therefore moves the rejected response one surface property closer to the chosen one, until only the meaning differs. The length gap is the score change from rewriting the rejected response to the length of the chosen one while preserving its meaning; the syntactic gap is the further change from rewriting



(a) (b)

Figure 1: (a) The chosen–rejected gap is decomposed along a meaning-preserving rewrite path that changes one property per step; a second rewrite under the same specification (twin) gives the scale ν against which gaps are read. (b) Directional semantic correctness C against length sensitivity L for the 25 models (bars: 95% CIs). Length-handling objectives move L both ways, but none raises C above plain DPO; only training on surface-matched pairs (rejected side rewritten to the chosen side's length, or length and sentence pattern) as ordinary pairs does.

the length-matched response into the sentence pattern of the chosen one; the semantic gap is what remains once length and syntax are controlled. Section 3 defines the three formally. The components are not perfectly disjoint, because length and syntax co-vary, so we make them well defined by a *residual* attribution: length is controlled first, and the length gap receives every change that comes with matching length, including the syntax that changes along with it. The convention is deliberately generous to length, so a finding that alignment is insensitive to the length gap is conservative.

To measure the three components we construct, for every TL;DR pair, two groups of meaning-preserving rewrites of the rejected response: y_1'', y_2'', with exactly the token count of the chosen response but the rejected response's sentence pattern (only length changes), and y_1', y_2', which additionally adopt the chosen response's sentence pattern (only syntax changes). Within each group the first rewrite lies on the decomposition path and the second, produced under the same specification and differing only in wording, is its twin; their discrepancy gives the scale of within-specification variability against which every gap is read. Dividing by this scale yields three metrics (Section 4): the unsigned semantic separation M, the syntactic sensitivity S and the length sensitivity L (closer to 1 is better for the last two); a value of 1 means that the channel moves the score no more than rewording within the specification does. Direction, which M does not carry, is measured by the fraction of pairs in which the model prefers the chosen meaning at matched length and sentence pattern, $\Pr(g_1 > 0)$, and by the signed semantic gap $\mathbb{E}\, g_1/\bar{\nu}$. Figure 1 summarizes the construction and the main finding.

Our experiments on 25 models of Qwen2.5-1.5B-Instruct and 16 models of Llama-3.2-3B-Instruct (Section 5) show the following. (i) The decomposition is reasonable but its components stay entangled: in controlled comparisons S and L move in opposite directions, mean-length penalties raise both, and sharp DRO reweighting compresses all three; only a residual attribution is well defined. (ii) Length penalties move L, in both directions, but no method at its usual strength moves S, and none improves directional semantic correctness or the signed semantic gap over plain DPO on the same pairs. (iii) Every method's S and L exceed 1, with S pinned at 1.35–1.63 for trained and untrained models alike, on both backbones. (iv) In a controlled comparison on identical data, training on the *surface-matched* pairs themselves, the pairs (y_w, y_1'') and (y_w, y_1') in which the rejected side is rewritten to the chosen side's length, or to its length and sentence pattern (Section 3.2), flattened into ordinary pairs, raises directional semantic correctness by 4–6 points, a gain that M does not register. (v) Together, (ii)–(iv) are what syntactic shadowing predicts: length penalties control length as a number and leave in place the syntactic channel that travels with it, whereas surface-matched training, which controls length as text, improves semantic preference; which surface property carries the gain is not decided.

In summary, our contributions are:

- a decomposition of the preference-alignment gap into length, residual syntactic and residual semantic components, with three scale-normalized metrics M, S, L and two directional outcomes;

- a benchmark of 1,979 TL;DR prompts with token-exact counterfactual rewrites and twins $(y_w, y_l, y_1'', y_2'', y_1', y_2')$, independently regenerated twins for 1,899 of them, and a disjoint training set of 1,998 prompts with the same rewrites;
- a sharp upper bound on the expected magnitude of the alignment gap in terms of the semantic and length gaps and a twin-calibrated syntactic distance (Sections A and 3.4);
- an audit of 25 models on one backbone and 16 on a second, showing that controlling length does not fix preference alignment: the semantic shortfall is not in the length channel, no method at its usual strength moves the syntactic channel, and surface-matched pairs improve directional semantics.

2 Related Work

Preference optimization learns from pairwise judgments either through a reward model and PPO (Schulman et al., 2017; Ouyang et al., 2022) or directly, as in DPO and its variants (Rafailov et al., 2023; Azar et al., 2024; Meng et al., 2024). Reward models, RLHF pipelines and LLM judges favor longer responses (Singhal et al., 2023; Dubois et al., 2024), and the remedies reviewed in Section 1 act on the objective, the data, the reward model or the RL update; all of them treat length as a confound to be removed.

Minimal pairs and counterfactual edits isolate one property of a text at a time, in evaluation (Warstadt et al., 2020; Ribeiro et al., 2020) and in training (Kaushik et al., 2020). Those edit an input to test a model's invariance or to flip a label; we edit one response of a preference pair, token-exactly and against a twin-calibrated scale, so that the model's own preference gap can be attributed to length, syntax and meaning. Work on likelihood displacement documents that DPO-style objectives can lower the likelihood of both responses of a pair while preserving their margin (Razin et al., 2025; Yuan et al., 2025).

3 Mathematical Formulation

This section makes the three components of Eq. (2) precise and bounds the expected magnitude of the alignment gap in terms of them; every symbol is defined once here and reused unchanged in Sections 4 and 5.

3.1 Setup

A preference pair $Z = (x, y_w, y_l)$ consists of a prompt x, a chosen response y_w and a rejected response y_l; as in standard preference optimization (Rafailov et al., 2023), pairs are drawn from a distribution $\mathcal{D}$, here the TL;DR pairs retained by the construction of Section 4.1. A response is a token sequence $y = (t_1, \ldots, t_{l(y)})$, and π_θ is the policy under study. We describe a response by three properties. Its *meaning* $m(y)$ is its propositional content given x: $m(y) = m(y')$ iff y and y' entail each other given x, the mutual-entailment criterion for paraphrase (Androutsopoulos & Malakasiotis, 2010). Its *sentence pattern* $s(y)$ is its constituency parse with the words removed, truncated to the top levels that record clause order, subordination and clause boundaries (a syntactic template; Iyyer et al., 2018). Its *length* $l(y) = |y|$ is its token count under the tokenizer of the evaluated model. Responses that agree in m, s and l can still differ in wording, which we treat as within-specification variability. The rewrites meet these specifications only up to the accuracy of the LLM writer and judges (Section B.3). The *score functional* is the sequence log-probability

$$\log \pi_\theta(y) \;=\; \sum_{i=1}^{l(y)} \log \pi_\theta(t_i \mid x, t_{<i}), \tag{3}$$

with the conditioning on x left implicit. The alignment gap of a pair is $G_\theta = \log \pi_\theta(y_w) - \log \pi_\theta(y_l)$ as in Eq. (1); the model prefers y_w iff $G_\theta > 0$. Each pair is extended to a *family* by the rewrites of Section 3.2, and $\mathbb{E}$ and $\Pr$ without subscript average over the families of the benchmark.

3.2 The rewrite path and the three residual gaps

The rewrites act on y_l alone: y_w is the target the pair is meant to teach and stays fixed, and the decomposition asks why y_l loses to it, for its length, for its sentence pattern, or for what it says.

Definition 1 (Length- and syntax-aligned rewrites)**.** Given a pair (x, y_w, y_l), a *length-aligned rewrite* of y_l is a response y_1'' with $m(y_1'') = m(y_l)$, $s(y_1'') = s(y_l)$ and $l(y_1'') = l(y_w)$, and a *syntax-aligned*

rewrite is a response y_1' with $m(y_1') = m(y_l)$, $s(y_1') = s(y_w)$ and $l(y_1') = l(y_w)$. The *twins* y_2'' and y_2' are second realizations under the same constraints as y_1'' and y_1'.

Both rewrites take their targets from y_w, so every gap below is a property of the pair Z; in the construction of Section 4.1, y_1' is moreover a rearrangement of the words of y_1'', which Section 3.4 uses.

The four responses form a path from the chosen to the rejected response along which each step changes exactly one property:

$$y_w \xrightarrow{\text{meaning}} y_1' \xrightarrow{\text{syntax}} y_1'' \xrightarrow{\text{length}} y_l. \tag{4}$$

We call (y_w, y_1'') the *length-matched* pair and (y_w, y_1') the *length- and syntax-matched* pair, and both *surface-matched pairs*: each keeps the chosen response of the original pair and replaces the rejected response by a rewrite that matches y_w on one or both surface properties, so that the two responses differ in meaning and sentence pattern but not in length, or in meaning alone.

Definition 2 (Residual gaps)**.** For a pair (x, y_w, y_l) with rewrites as in Definition 1,

$$g_1 = \log \pi_\theta(y_w) - \log \pi_\theta(y_1'), \quad g_2 = \log \pi_\theta(y_1') - \log \pi_\theta(y_1''), \quad g_3 = \log \pi_\theta(y_1'') - \log \pi_\theta(y_l) \tag{5}$$

are the *residual semantic gap*, the *residual syntactic gap* and the *length gap*.

The gaps telescope, $G_\theta = g_1 + g_2 + g_3$, for any rewrites satisfying Definition 1. Across pairs the three gaps take both signs, so the magnitude of the alignment gap is bounded by theirs,

$$\mathbb{E}\,|G_\theta| \;\le\; \mathbb{E}\,|g_1| + \mathbb{E}\,|g_2| + \mathbb{E}\,|g_3|, \tag{6}$$

with equality only when the three gaps have the same sign in every pair; this is Eq. (2) made precise.

The attribution is residual by construction. Because s is coarse, matching length while keeping the sentence pattern still forces small structural adjustments; the length step $y_l \to y_1''$ is taken first, so g_3 receives them together with everything else that comes with matching length, g_2 is the residual syntactic effect at fixed length, and g_1 the residual once both surface properties match those of y_w. Reversing the order would move the shared part of the two surface effects from g_3 to g_2 without changing their sum; we fix the length-first order because it favors length, and Section 5 argues against length.

3.3 Twin-discrepancy scale and metrics

Two responses that share meaning, sentence pattern and token count still differ in wording, and a fluent language model necessarily scores them differently. The twins of Definition 1 measure this difference; for a pair we define the *twin-discrepancy scale*

$$\nu \;=\; \tfrac{1}{2}\Big(\big|\log \pi_\theta(y_1') - \log \pi_\theta(y_2')\big| + \big|\log \pi_\theta(y_1'') - \log \pi_\theta(y_2'')\big|\Big), \tag{7}$$

and write $\bar{\nu}$ for its mean over the benchmark. It is a ruler rather than a target, the size of a score difference between two responses that satisfy the same specification, and every metric below is a ratio to it (Table 10 gives both terms). It is not a zero-mean noise: the two rewrites of a family are produced by the same writer call in a fixed order and revised together by the auditor (Section 4.1), and the first version scores systematically higher than the second under every model we evaluated (by 4.7 and 4.9 on average for the untrained model, in 66–69% of families). We therefore read ν as the within-specification variability of the rewrite pipeline and check in Section 5.4 how the results change when the twin roles are swapped; the conditionally i.i.d. twins assumed in Section 3.4 are an idealization of this pipeline.

One further step is needed for g_3 alone, because y_1'' and y_l are the only two responses on the path with different token counts. Writing $\mu_a = \log \pi_\theta(a)/l(a)$ for the average per-token log-probability of a response a, the score difference of two responses splits exactly into a content part and a mechanical part,

$$\log \pi_\theta(a) - \log \pi_\theta(b) \;=\; \underbrace{\bar{l}\,(\mu_a - \mu_b)}_{\text{content}} \;+\; \underbrace{\big(l(a) - l(b)\big)\,\bar{\mu}}_{\text{mechanical}}, \qquad \bar{l} = \tfrac{l(a)+l(b)}{2}, \;\; \bar{\mu} = \tfrac{\mu_a+\mu_b}{2}, \tag{8}$$

where the mechanical part is the cost of emitting more tokens irrespective of what they say. It vanishes for g_1, g_2 and ν; for g_3 we keep only the content part, $g_3^{\mathrm{c}} = \bar{l}\,(\mu_{y_1''} - \mu_{y_l})$, so that length sensitivity measures how the model scores what the extra tokens say, not that there are extra tokens; the token-count term that penalties act on is excluded by design (Section A.6 keeps the raw gap).

Definition 3 (Unsigned semantic separation, syntactic sensitivity, length sensitivity)**.** With $\bar{\nu} = \mathbb{E}\,\nu$,

$$M = \frac{\mathbb{E}|g_1|}{\bar{\nu}}, \qquad S = \frac{\mathbb{E}|g_2|}{\bar{\nu}}, \qquad L = \frac{\mathbb{E}|g_3^{\mathrm{c}}|}{\bar{\nu}}. \tag{9}$$

Definition 4 (Directional semantic outcomes)**.** The *directional semantic correctness* is $C = \Pr(g_1 > 0)$, the fraction of families in which the model prefers the chosen summary over the rejected meaning at matched length and sentence pattern, and the *signed semantic gap* is $\bar{g}_1 = \mathbb{E}\,g_1/\bar{\nu}$.

All five quantities are in units of ν. M is a separation, not a margin (a model that strongly prefers the rejected meaning also has a large M); direction is carried by C and $\bar{g}_1$, with which every claim about semantic alignment is made. The three sensitivities use absolute values because the signed means $\mathbb{E}\,g_2$ and $\mathbb{E}\,g_3$ measure *bias*, whether the model prefers one sentence pattern or one length on average, and can vanish while per-example effects are large; $\mathbb{E}|g|$ measures the size of the per-example effect, which is what enters the bound of Section 3.4. Two decision-level quantities complete the picture: the pairwise accuracy $\mathrm{acc} = \Pr(G_\theta > 0)$ on the original pair, and the likelihood displacement $\Delta_{\mathrm{disp}} = \mathbb{E}\,\log \pi_\theta(y_w) - \mathbb{E}\log \pi_{\mathrm{ref}}(y_w)$ of the chosen responses relative to the untrained reference, which flags *large displacement* (Section 4.2).

3.4 Decomposition inequality

How much each component matters is a question about magnitudes across pairs (Eq. (6)). Consider an auxiliary law that draws $Z \sim \mathcal{D}$ and then fresh rewrites satisfying Definition 1, without joint family filtering: the length twins are conditionally i.i.d. given Z, and the syntax twins are conditionally i.i.d. and independent of y_2'' given (Z, y_1''). Denote its expectation by $\mathbb{E}_0$ and set $\bar{\nu}_0 = \mathbb{E}_0\nu$. With $A_i = \log \pi_\theta(y_i')$ and $B_i = \log \pi_\theta(y_i'')$, define

$$a = \mathbb{E}_0|g_1| + \mathbb{E}_0|g_3|, \qquad d = \frac{1}{2}\mathbb{E}_0\left[\sum_{i=1}^{2}|A_i - B_2|\right] - \bar{\nu}_0. \tag{10}$$

Here g_3 retains the mechanical term of Eq. (8); the proof is in Section A.

Theorem 1 (Alignment-gap upper bound)**.** *Assume integrable scores under the auxiliary sampling law. Then $d \geq 0$ and*

$$\mathbb{E}_{\mathcal{D}}|G_\theta| \leq \min\left\{\sum_{j=1}^{3}\mathbb{E}_0|g_j|, \quad a + \frac{d + \sqrt{d^2 + 4ad}}{2}\right\}. \tag{11}$$

The second term is optimal over admissible score laws when only (a, d) are fixed.

Remark. By the energy-distance identity (Bellemare et al., 2017) applied conditionally under the auxiliary law, d is the expected Cramér distance between the conditional score laws of the syntax- and the length-aligned rewrite, the syntactic discrepancy net of twin variability; $d \geq 0$ is thus the cross-twin, population counterpart of $S \geq 1$ (Section 5.4). The syntactic channel enters the second term only through d. When $d = 0$, the syntax-aligned rewrite is scored as if it were a further length-aligned rewrite and $\mathbb{E}_{\mathcal{D}}|G_\theta| \leq a$: the alignment gap is carried by semantics and length alone. Otherwise the term lies between $a + d$ and $2a + d$ and equals $a + \sqrt{ad} + d/2 + O(d^{3/2}a^{-1/2})$ for $d \ll a$, so the channels interact through a geometric mean rather than add; it improves on the cross-path triangle bound $a + d + \bar{\nu}_0$ iff $d(a - \bar{\nu}_0) < \bar{\nu}_0^2$ (Theorem 2 fixes $\bar{\nu}_0$ as well). The bound concerns magnitude, not sign, d also absorbs variation induced by the parent rewrite y_1'', and the auxiliary law idealizes a pipeline whose twins are co-written and jointly filtered, so on the benchmark a and d are diagnostics only: $\hat{d} > 0$ for every model, the first term is the tighter one, and $\mathbb{E}|G_\theta|$ is 0.44–0.57 of it, about half of the component magnitudes cancelling within pairs (Section A.6).

4 Benchmark and Metrics

Section 3 assumed that rewrites with the properties of Definition 1 exist for every pair. We describe how we construct them, then turn the metrics of Definition 3 into an evaluation protocol.

4.1 Constructing y_1'', y_2'' and y_1', y_2'

We start from the TL;DR summarization preference data (Völske et al., 2017; Stiennon et al., 2020): a prompt x is a Reddit post and y_w, y_l are two human-written summaries, one preferred by annotators. We sample 2,000 source pairs; each becomes a *family* consisting of the prompt, the two human

Table 1: A benchmark family (prompt omitted; token counts in parentheses). y_2'' and y_2', the twins of y_1'' and y_1' respectively, are not shown; each differs from its path rewrite only in wording.

y_w	I work with a guy who's 100% illiterate, and I'm trying to figure out how he does it. (27)
y_l	Reddit: how does a guy with no reading or writing skills survive in today's society? (18)
y_1''	Reddit: how exactly does a guy with absolutely no reading or writing skills at all manage to survive and function in today's modern society? (27)
y_1'	Reddit, a guy has absolutely no reading or writing skills, and I'm asking how he survives and functions in today's modern society. (27)

summaries and four rewrites, specified in two steps, written by an LLM writer and checked by an adversarial auditor (Section B.1). The first step aligns length: the rejected summary is rewritten to *exactly* the token count of the chosen summary, keeping its meaning and its own sentence pattern, in two versions y_1'' and y_2''. The second step aligns syntax: y_1'' is rearranged into the sentence pattern of the chosen summary at fixed token count, a structural rearrangement of the same words rather than a fresh paraphrase, again in two versions y_1' and y_2' (Table 1). Subscript 1 marks the path rewrite and 2 its twin; all four come from one writer call (Section 3.3).

An assembler drops every family that violates a hard constraint on token counts or identity (Section B.2); 1,979 of the 2,000 families pass, all with distinct prompts. All rewrites are on the rejected side; the chosen summary is always the human original (Section 6). A disjoint training set of 1,998 prompts is used for the counterfactual-data controls of Section 5.

4.2 Evaluation protocol

We evaluate a model by scoring the six responses of every family with Eq. (3), forming the per-family gaps Eq. (5) and twin-discrepancy scale Eq. (7), and reporting the metrics of Definitions 3 and 4 together with acc and Δ_{disp}. Each scale-normalized metric is a ratio to the model's own $\bar{\nu}$, which makes models with different score scales comparable ($M = 3$: the meanings are three rewording discrepancies apart). Comparisons between models are made on the same families, with paired bootstrap differences. The displacement Δ_{disp} is needed because DPO-type objectives constrain only the relative margin: a policy can raise it by lowering both likelihoods, and such a collapse rescales every gap and the twin-discrepancy scale at once. We flag *large displacement* when $\Delta_{\text{disp}} < -60$ (plain DPO sits near -15); the threshold is a judgment call, and such models are included throughout (Section C.5).

5 Experiments

We audit 25 models on Qwen2.5-1.5B-Instruct and 16 on Llama-3.2-3B-Instruct: are the three components separate channels (Section 5.2), which channel do length-handling objectives move and does that help meaning (Section 5.3), and how large is each channel relative to rewording (Section 5.4)?

5.1 Setup

Models. The 25 main models share the backbone Qwen2.5-1.5B-Instruct (Yang et al., 2024) and are scored on the 1,979 families of Section 4; Base, the untrained instruct model, is also the reference and the initialization of every trained model. Ten models reproduce SOTA methods on the 92,858-pair TL;DR preference split (Table 7): DPO, Robust-DPO (Chowdhury et al., 2024), Dr. DPO and the seven length-handling methods of Section 1. Eight in-house Dr. DPO variants on the same data vary its reweighting, add the R-DPO regularizer, or add a *length-invariance penalty*: λ times the squared difference between the mean alignment gap G_θ of the chosen-longer pairs and that of the remaining pairs of a micro-batch, at $\lambda = 5 \cdot 10^{-4}$ (I4) and $2.5 \cdot 10^{-4}$ (I4b), a REx-type invariance penalty with the two length groups as environments (Krueger et al., 2021). Six *counterfactual-data controls* use the rewrites themselves as training data, to test whether surface-matched pairs (Section 3.2) rather than a length penalty improve semantic alignment: on the disjoint 1,998-prompt set with its own rewrites, the arms share prompts and seeds and differ only in which pairs enter the loss, from Dr. DPO on the original pairs, through arms that add the length- and syntax-matched views, to two *flattened* arms in which every view is an ordinary pair (Section C); they see $46\times$ less data than the other groups, so only within-group differences count. The whole protocol is repeated on Llama-3.2-3B-Instruct with the same configurations (16 models; Section F.5).

Table 2: Metrics of the 25 models (1,979 families; Section 4.2). $\to 1$: S and L are ratios to the twin-discrepancy scale $\bar{\nu}$; closer to 1, the rewording floor, is better. °Reference-free. †Large displacement ($\Delta_{\text{disp}} < -60$; Section C.5). Controls: three-seed means on 46× less data. Bold: discussed in text; paired CIs in Table 8 (Section D).

Method	Length handling	acc ↑	M ↑	$S \to 1$	$L \to 1$	C ↑	$\bar{g}_1$ ↑	Δ_{disp}
BASE (untrained)	–	0.518	3.238	1.521	1.490	0.774	2.36	0.0
SOTA methods (92,858 training pairs)								
DPO	none	0.580	3.535	1.508	1.551	0.790	2.73	−14.7
Robust-DPO	none	0.568	3.531	1.520	1.536	0.796	2.73	−11.5
Dr. DPO	pair reweighting (DRO)	0.592	3.457	1.498	1.538	0.790	2.70	−27.6
R-DPO	explicit length regularizer	0.565	3.532	1.539	**1.661**	0.793	2.74	−5.8
LD-DPO	length desensitization	0.548	3.495	1.506	1.492	0.794	2.69	−11.0
SamPO	token down-sampling	0.547	3.456	1.501	1.469	0.792	2.64	−8.3
SimPO°	avg. log-prob reward	0.511	3.329	1.546	**1.362**	0.771	2.44	−0.5
ORPO°	avg. log-prob odds ratio	0.514	3.331	1.495	1.486	0.784	2.52	+22.6
Tie training	tie pairs (data)	0.543	3.426	1.491	1.486	0.789	2.63	−1.6
LMPO°†	length-controlled margin	0.531	3.638	1.577	1.208	0.754	2.56	−87.4
Dr. DPO variants (same training pairs)								
β'=0.25†	pair reweighting	0.587	3.237	1.413	1.213	0.766	2.36	−294.7
β'=0.5†	pair reweighting	0.604	3.231	1.441	1.290	0.774	2.39	−132.9
β'=2	pair reweighting	0.579	3.552	1.510	1.548	0.792	2.74	−13.8
length-stratified	pair reweighting	0.589	3.506	1.479	1.554	0.797	2.73	−32.7
buffered (128 losses)	pair reweighting	0.596	3.418	1.493	1.544	0.794	2.66	−31.3
+ R-DPO regularizer	explicit length regularizer	0.596	3.472	1.520	1.654	0.793	2.70	−33.0
+ $\lambda = 5 \cdot 10^{-4}$ (I4)	mean-length-bias penalty	0.581	3.434	**1.630**	**2.618**	0.795	2.68	−9.9
+ $\lambda = 2.5 \cdot 10^{-4}$ (I4b)	mean-length-bias penalty	0.589	3.484	**1.618**	**2.330**	0.793	2.71	−10.1
Counterfactual-data controls (1,998 prompts, 3 seeds)								
Dr. DPO, original pairs	–	0.519	3.279	1.499	1.479	0.775	2.42	−0.3
DPO, KL-DRO over views	rewrites as views	0.520	3.273	1.485	1.489	0.781	2.43	+0.4
Dr. DPO, KL-DRO over views	rewrites as views	0.520	3.277	1.479	1.487	0.781	2.44	+0.1
Dr. DPO, uniform views	rewrites as views	0.531	3.205	**1.393**	1.521	0.796	2.49	−3.3
flattened pairs, 3× steps	rewrites as pairs	0.545	3.203	**1.357**	1.571	**0.834**	2.72	−47.9
flattened pairs, step-matched	rewrites as pairs	0.545	3.182	**1.353**	1.556	**0.820**	2.63	−25.5

5.2 THE THREE COMPONENTS ARE NOT SEPARATE CHANNELS

The gaps add up pair by pair and Eq. (6) bounds the alignment gap by their magnitudes, but neither makes them separate channels: the two surface components move together, differently in each training regime (Table 2). Within the controls they trade off ($\text{corr}(S, L) = -0.97$ over the six arms); under a mean-length penalty they rise together, the I4 variants having both the largest L (2.33, 2.62) and the largest S (1.62, 1.63); under sharp DRO reweighting ($\beta' \leq 0.5$) both fall (S 1.41–1.44, L 1.21–1.29) while the twin scale grows fourfold or more. The three metrics are therefore reported separately, with Δ_{disp}, under the length-favoring residual attribution of Section 3.2.

5.3 LENGTH PENALTIES MOVE L AND NOTHING ELSE

Table 2 reports all 25 models (Fig. 1(b) plots C against L); four findings stand out.

No method moves the syntactic channel. The nine SOTA methods other than LMPO, the six that handle length included, keep S within 0.03 of the untrained 1.521 (1.491–1.546; SD 0.019 over the 14 same-data models without large displacement or a mean-length penalty); only a fourfold increase of R-DPO's penalty moves it, to 1.63 (Section F.2). These are ratios to each model's $\bar{\nu}$; Table 10 gives the absolute magnitudes.

The length channel moves, in both directions. Across models L varies 4.5 times as much as S (SD 0.298 vs. 0.067). SimPO lowers it to 1.362, the lowest value but LMPO's; R-DPO, the most explicit length regularizer, raises it above plain DPO (1.661 vs. 1.551), and the I4 variants to 2.330 and 2.618. I4 drives the penalized mean length bias to its minimum yet has the largest per-example sensitivity: a statistic whose terms cancel can be minimized without desensitizing the model.

Moving L does not improve semantic alignment. At matched length and sentence pattern, DPO prefers the chosen meaning in 79.0% of families against 77.4% for BASE, a small gain (+0.016 [+0.002, +0.031]). No length-handling objective improves on DPO in the principal analysis (Fig. 1(b); Table 8): the intervals cap any gain in C at 0.014 and in $\bar{g}_1$ at 0.06, and SimPO, SamPO, ORPO and tie training, which push L below BASE, are those whose signed gap falls significantly below DPO's (by 0.09–0.29; SimPO also in C). Under the swapped twin designation alone, R-DPO gains 0.012 in C, which the judge filters remove (Table 12 and Section B.3). A sixteenfold sweep of R-DPO's penalty moves L from 1.55 to 2.32 while C stays within 0.007 of DPO (Section F.2).

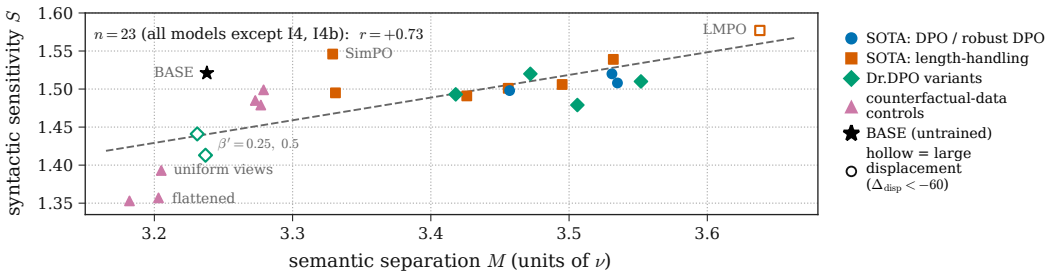


Figure 2: S against M for 23 models, all but I4 and I4b, whose mean-length penalty raises S alone (markers as in Fig. 1(b); dashed: OLS): $r = +0.73$, 95% CI $[0.47, 0.89]$.

Surface-matched pairs help when they enter training as ordinary pairs. On identical prompts, the two flattened arms reach $C = 0.820$ (step-matched) and 0.834 ($3\times$ steps), 4–6 points above Dr. DPO on the original pairs and with the rewrites as KL-DRO views, with signed gaps larger by 0.19–0.30 (Table 8) and acc higher by 0.026; in the KL-DRO arms the same rewrites leave C almost unchanged (0.781 vs. 0.775), and length-matched pairs alone gain as much as length- and syntax-matched ones (Section F.3), so the gain does not require explicit syntax matching. M misses the gain: it is lower for the flattened arms (3.182–3.203 vs. 3.279); their lower S is a change of ratio, $\mathbb{E}|g_2|$ in raw log-probability rising from 12.9 to 21–27 while $\bar{\nu}$ roughly doubles (Table 10).

5.4 Syntax stays at a fixed distance above the twin-discrepancy scale

Two facts about scale follow from Table 2. First, every S and L exceeds 1 (Eq. (7)): changing the sentence pattern or the length of a rejected summary moves every model's score more than rewording does, trained or not. Second, syntax sits at a fixed distance above this scale, about one and a half times it before and after training: $S = 1.521$ for BASE and 1.353–1.630 for all trained models; on Llama-3.2-3B-Instruct the untrained value is 1.520 and the trained ones 1.32–1.53. Both facts survive changes to the ruler: swapping the roles of the ordered twins shifts C by 0.03–0.05 but moves S by at most 0.025 and leaves DPO $>$ BASE, SimPO $<$ DPO and the flattened-arm gains significant (Section D); and regenerating the twins as independent draws or filtering the families by LLM judges lowers S by up to 0.2 but keeps it above 1 for every model (Section F).

An association with syntax. Across the 23 models other than the two I4 variants, M rises with S ($r = +0.73$; Fig. 2). Within models, the semantic gap co-moves with the syntactic residual more than with the length residual in 22 of 25 models, significantly in 17, and already in BASE (Section D); original-pair accuracy follows L ($\rho_s = 0.71$), not S (0.20).

6 Discussion and Conclusion

On two backbones, length penalties move L but no length-handling objective raises C above plain DPO; no method at its usual strength moves S, which stays about 1.5 twin-discrepancy units above rewording as before training; what improves directional semantics is training on flattened surface-matched pairs. This is what syntactic shadowing predicts: a syntactic component that co-varies with length is left in place by controlling length as a number and reached only by matching length as text, even under the residual attribution that credits length with every syntactic change travelling with it. Whether the gain needs the surface match at all, rather than any rewritten rejected summary, a control trained on unconstrained paraphrases would decide, and the per-family associations (Section 5.4) are weak and present before training, so the mechanism remains a hypothesis that the decomposition makes testable. In practice, report surface bias as sensitivities with $\bar{\nu}$ and Δ_{disp}, not signed means, and semantic alignment directionally.

Limitations. C ranks constructed responses under the original pair's label, assumed to survive the rewrite, and is not a validated judgment of generated summaries. The rewrites are LLM-generated, screened by LLM judges that agree poorly with each other (Section B.3), and exist only on the rejected side, so a model could partly separate human from LLM text rather than meaning from form; chosen-side rewrites with blinded human checks would address both. Some sensitivity to sentence structure may be legitimate, the evidence covers one domain, and DPO uses a smaller β than its variants. This paper diagnoses; it does not prescribe.

AI USE STATEMENT

In this work, we used generative AI tools for tasks with required disclosure: generating the synthetic benchmark, whose rewrites and twins were written and audited by LLM agents (Sections B.1 and F.4) and screened by LLM judges (Section B.3); implementing the training, evaluation and analysis code. For tasks with recommended disclosure, we used them to edit code, write figure scripts, draft and edit parts of the paper, and identify and format references. Every rewrite passed code-enforced constraints (Section B.2), judge screening is reported as a robustness check rather than as validation, and every reported number was computed by scripts from the per-family score files. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The anonymous repository (Footnote 1) contains the training, evaluation and benchmark-construction code, the benchmark, the 1,998-prompt training set and the per-family scores of the 37 checkpoints behind Table 2, from which the table can be recomputed. The construction is specified in Section B, the metrics in Sections 3 and 4.2, the configuration of every run in Table 7, and the splits and bootstrap protocol in Section B.4.

REFERENCES

Ion Androutsopoulos and Prodromos Malakasiotis. A survey of paraphrasing and textual entailment methods. *J. Artif. Intell. Res.*, 38:135–187, 2010. doi: 10.1613/JAIR.2985. URL https://doi.org/10.1613/jair.2985.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Rémi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *International Conference on Artificial Intelligence and Statistics, 2-4 May 2024, Palau de Congressos, Valencia, Spain*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4447–4455. PMLR, 2024. URL https://proceedings.mlr.press/v238/gheshlaghi-azar24a.html.

Marc G. Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The cramer distance as a solution to biased wasserstein gradients. *CoRR*, abs/1705.10743, 2017. URL http://arxiv.org/abs/1705.10743.

RALPH ALLAN BRADLEY and MILTON E. TERRY. Rank analysis of incomplete block designs: The method of paired comparisons. *Biometrika*, 39(3-4):324–345, 12 1952. ISSN 0006-3444. doi: 10.1093/biomet/39.3-4.324. URL https://doi.org/10.1093/biomet/39.3-4.324.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with GPT-4. *CoRR*, abs/2303.12712, 2023. doi: 10.48550/ARXIV.2303.12712. URL https://doi.org/10.48550/arXiv.2303.12712.

Lichang Chen, Chen Zhu, Jiuhai Chen, Davit Soselia, Tianyi Zhou, Tom Goldstein, Heng Huang, Mohammad Shoeybi, and Bryan Catanzaro. ODIN: disentangled reward mitigates hacking in RLHF. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7935–7952. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/chen24bn.html.

Sayak Ray Chowdhury, Anush Kini, and Nagarajan Natarajan. Provably robust DPO: Aligning language models with noisy feedback. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=yhpDKSw7yA.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 4299–4307, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *CoRR*, abs/2404.04475, 2024. doi: 10.48550/ARXIV.2404.04475. URL https://doi.org/10.48550/arXiv.2404.04475.

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

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Jiwoo Hong, Noah Lee, and James Thorne. ORPO: monolithic preference optimization without reference model. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 11170–11189. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.626. URL https://doi.org/10.18653/v1/2024.emnlp-main.626.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. Adversarial example generation with syntactically controlled paraphrase networks. In Marilyn A. Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pp. 1875–1885. Association for Computational Linguistics, 2018. doi: 10.18653/V1/N18-1170. URL https://doi.org/10.18653/v1/n18-1170.

Divyansh Kaushik, Eduard H. Hovy, and Zachary Chase Lipton. Learning the difference that makes A difference with counterfactually-augmented data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=Sklgs0NFvr.

David Krueger, Ethan Caballero, Jörn-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Rémi Le Priol, and Aaron C. Courville. Out-of-distribution generalization via risk extrapolation (rex). In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5815–5826. PMLR, 2021. URL http://proceedings.mlr.press/v139/krueger21a.html.

Gengxu Li, Tingyu Xia, Yi Chang, and Yuan Wu. Length-controlled margin-based preference optimization without reference model. *arXiv preprint arXiv:2502.14643*, 2025.

Tianqi Liu, Wei Xiong, Jie Ren, Lichang Chen, Junru Wu, Rishabh Joshi, Yang Gao, Jiaming Shen, Zhen Qin, Tianhe Yu, Daniel Sohn, Anastasia Makarova, Jeremiah Zhe Liu, Yuan Liu, Bilal Piot, Abe Ittycheriah, Aviral Kumar, and Mohammad Saleh. RRM: Robust reward model training mitigates reward hacking. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=88AS5MQnmC.

Wei Liu, Yang Bai, Chengcheng Han, Rongxiang Weng, Jun Xu, Xuezhi Cao, Jingang Wang, and Xunliang Cai. Length desensitization in direct preference optimization, 2024. URL https://arxiv.org/abs/2409.06411.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. In *Second Conference on Language Modeling*, 2025b. URL https://openreview.net/forum?id=5PAF7PAY2Y.

Junru Lu, Jiazheng Li, Siyu An, Meng Zhao, Yulan He, Di Yin, and Xing Sun. Eliminating biased length reliance of direct preference optimization via down-sampled KL divergence. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 1047–1067. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.60. URL https://doi.org/10.18653/v1/2024.emnlp-main.60.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/e099c1c9699814af0be873a175361713-Abstract-Conference.html.

Christian Moya, Alex Semendinger, Guang Lin, and Elliott Thornley. Spurious correlation learning in preference optimization: Mechanisms, consequences, and mitigation via tie training. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Hpfybj9wkd.

Long Ouyang, Jeffrey 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 F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, volume ACL 2024 of *Findings of ACL*, pp. 4998–5017. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.297. URL https://doi.org/10.18653/v1/2024.findings-acl.297.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Noam Razin, Sadhika Malladi, Adithya Bhaskar, Danqi Chen, Sanjeev Arora, and Boris Hanin. Unintentional unalignment: Likelihood displacement in direct preference optimization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=uaMSBJDnRv.

Marco Túlio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with checklist. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 4902–4912. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ACL-MAIN.442. URL https://doi.org/10.18653/v1/2020.acl-main.442.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL http://arxiv.org/abs/1707.06347.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in RLHF. *CoRR*, abs/2310.03716, 2023. doi: 10.48550/ARXIV.2310.03716. URL https://doi.org/10.48550/arXiv.2310.03716.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize with human feedback. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1f89885d556929e98d3ef9b86448f951-Abstract.html.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. Tl;dr: Mining reddit to learn automatic summarization. In Lu Wang, Jackie Chi Kit Cheung, Giuseppe Carenini, and Fei Liu (eds.), *Proceedings of the Workshop on New Frontiers in Summarization, NFiS@EMNLP 2017, Copenhagen, Denmark, September 7, 2017*, pp. 59–63. Association for Computational Linguistics, 2017. doi: 10.18653/V1/W17-4508. URL https://doi.org/10.18653/v1/w17-4508.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020. doi: 10.1162/tacl_a_00321. URL https://aclanthology.org/2020.tacl-1.25/.

Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jiawei Chen, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. Towards robust alignment of language models: Distributionally robustifying direct preference optimization. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=CbfsKHiWEn.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 55204–55224. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/xu24t.html.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Hui Yuan, Yifan Zeng, Yue Wu, Huazheng Wang, Mengdi Wang, and Liu Leqi. A common pitfall of margin-based language model alignment: Gradient entanglement. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=YaBiGjuDiC.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. *CoRR*, abs/2507.18071, 2025. doi: 10.48550/ARXIV.2507.18071. URL https://doi.org/10.48550/arXiv.2507.18071.

A Proofs for the alignment-gap bound

Let P_0 denote the auxiliary law in Section 3.4. All rewrite distributions and conditional expectations below refer to P_0. It draws Z from $\mathcal{D}$, but imposes no joint selection on fresh rewrite families.

Consequently,

$$\mathbb{E}_0|G_\theta| = \mathbb{E}_{\mathcal{D}}|G_\theta|. \tag{12}$$

The quantities $\bar{\nu}_0$, a, and d must all be evaluated under P_0; $\bar{\nu}_0$ need not equal the benchmark noise scale $\bar{\nu}$ of Section 3.3. Denote the nonlinear bound by

$$\Phi(a, d) = a + \frac{d + \sqrt{d^2 + 4ad}}{2}. \tag{13}$$

Fix θ and write $w = \log \pi_\theta(y_w)$, $\ell = \log \pi_\theta(y_l)$, $H = \mathbb{E}_0|w - \ell|$, and $C = (Z, y_1'')$. Let F_C be the conditional CDF of A_1 given C, and let Q_Z be the conditional CDF of B_2 given Z. Separate the two contributions to the noise floor as

$$\nu_A = \frac{1}{2}\mathbb{E}_0|A_1 - A_2|, \qquad \nu_B = \frac{1}{2}\mathbb{E}_0|B_1 - B_2|, \qquad \bar{\nu}_0 = \nu_A + \nu_B. \tag{14}$$

A.1 Calibration by the twins

For independent real random variables X, Y and independent copies X', Y', the energy-distance identity (Bellemare et al., 2017) gives

$$\mathbb{E}|X - Y| - \frac{1}{2}\mathbb{E}|X - X'| - \frac{1}{2}\mathbb{E}|Y - Y'| = \int_{\mathbb{R}} (F_X(t) - F_Y(t))^2 \, dt. \tag{15}$$

Applying this identity conditionally and averaging yields

$$d = \mathbb{E}_{0,C} \int_{\mathbb{R}} (F_C(t) - Q_Z(t))^2 \, dt \geq 0. \tag{16}$$

Here the within-Q_Z term is ν_B after averaging, since the length twins are i.i.d. given Z. It is not computed from the conditional law of the fixed parent score B_1 given C.

For completeness, writing $\overline{F}_Z = \mathbb{E}_0[F_C \mid Z]$ gives

$$d = \mathbb{E}_0 \int_{\mathbb{R}} (\overline{F}_Z - Q_Z)^2 \, dt + \mathbb{E}_0 \int_{\mathbb{R}} (F_C - \overline{F}_Z)^2 \, dt. \tag{17}$$

Thus d includes both a marginal score-distribution discrepancy and variation induced by the parent rewrite. This interpretation does not require d to be a pure syntactic effect.

A.2 Proof of the alignment-gap upper bound

Fix C, let I be the interval between w and ℓ, and put $h = |w - \ell|$. Let A^I, B^I be the projections of A_1, B_2 onto I. Projection decreases their distances to the corresponding endpoints and preserves their CDFs on the interior of I. Reflecting the scores if necessary, take $w \geq \ell$. With

$$a_C = \mathbb{E}_0[|w - A_1| \mid C] + \mathbb{E}_0[|B_2 - \ell| \mid Z], \qquad d_C = \int_{\mathbb{R}} (F_C - Q_Z)^2 \, dt,$$

we obtain

$$\begin{aligned} h &= \big(w - \mathbb{E}_0[A^I \mid C]\big) + \big(\mathbb{E}_0[A^I \mid C] - \mathbb{E}_0[B^I \mid C]\big) + \big(\mathbb{E}_0[B^I \mid C] - \ell\big) \\ &\leq a_C + \int_I (Q_Z - F_C) \, dt \leq a_C + \sqrt{h \, d_C}. \end{aligned} \tag{18}$$

Taking expectations and applying Cauchy–Schwarz gives $H \leq a + \sqrt{Hd}$. Solving for H proves $H \leq \Phi(a, d)$.

The original path gives $H \leq a + \mathbb{E}_0|g_2|$. Together with Eq. (12), this proves Eq. (11).

To establish optimality with (a, d) fixed, let $h = \Phi(a, d) > 0$, $\delta = \sqrt{d/h}$, and take endpoints $w = h$, $\ell = 0$. Draw all rewrites independently, with scores supported on $\{0, h\}$ and

$$\Pr(A_i = h) = \frac{1 + \delta}{2}, \qquad \Pr(B_i = h) = \frac{1 - \delta}{2}. \tag{19}$$

The endpoint cost is $h(1 - \delta) = a$, the squared CDF discrepancy is $h\delta^2 = d$, and $H = h$. For $a = d = 0$, take all scores to be zero. This proves the sharpness claim in Theorem 1. Optimality ranges over preference-pair distributions and integrable score laws; it does not assert attainability for a fixed model on a fixed benchmark. The construction may change $\bar{\nu}_0$ and $\mathbb{E}_0|g_2|$, which are not fixed in this claim.

A.3 Refinement using the noise floor

When $\bar{\nu}_0$ is retained separately, the following result gives the optimal upper bound based on all three aggregate summaries.

Theorem 2 (Sharp bound with three fixed summaries)**.** *Under the assumptions of Theorem 1, $a \geq \bar{\nu}_0$ and*

$$\Psi(a,d,v) = \min\Big\{a + d + v,\ \Phi(a,d),\ \big(\sqrt{2(a-v)} + \sqrt{d}\big)^2\Big\}, \tag{20}$$

$$\mathbb{E}_0|G_\theta| \leq \Psi(a,d,\bar{\nu}_0). \tag{21}$$

Let $\mathfrak{P}(a,d,v)$ be the class of integrable score laws satisfying these conditional sampling assumptions and having summaries $(a,d,\bar{\nu}_0) = (a,d,v)$, with the preference-pair distribution allowed to vary. For every $a > v \geq 0$ and $d \geq 0$,

$$\sup_{P \in \mathfrak{P}(a,d,v)} \mathbb{E}_P|G_\theta| = \Psi(a,d,v). \tag{22}$$

The same equality holds for $a = v = 0$, with value d.

Cross-path bound. The length twins have the same law given Z, so the cross path gives

$$H \leq \mathbb{E}_0|w - A_1| + \mathbb{E}_0|A_1 - B_2| + \mathbb{E}_0|B_2 - \ell| = a + d + \bar{\nu}_0. \tag{23}$$

Bound using the noise floor. Let $J_w(t) = \mathbf{1}\{w \leq t\}$ and $J_\ell(t) = \mathbf{1}\{\ell \leq t\}$. Use the norm of $L^2(dP_{0,C}\,dt)$. The point-mass case of Eq. (15) gives

$$\|J_w - F_C\|_2^2 = \mathbb{E}_0|g_1| - \nu_A, \qquad \|Q_Z - J_\ell\|_2^2 = \mathbb{E}_0|g_3| - \nu_B. \tag{24}$$

Their sum is $a - \bar{\nu}_0 \geq 0$. Since $\|J_w - J_\ell\|_2^2 = H$ and $\|F_C - Q_Z\|_2^2 = d$, the triangle inequality yields

$$\begin{aligned}\sqrt{H} &\leq \|J_w - F_C\|_2 + \|F_C - Q_Z\|_2 + \|Q_Z - J_\ell\|_2 \\ &\leq \sqrt{2(a-\bar{\nu}_0)} + \sqrt{d}.\end{aligned} \tag{25}$$

The minimum of the three branches proves Eq. (21). If $a = \bar{\nu}_0$, both endpoint norms vanish. The conditional score laws are then point masses at their respective endpoints, so $a = \bar{\nu}_0 = 0$.

A.4 Sharpness with three fixed summaries

We construct score laws with precisely the prescribed summaries (a,d,v). All rewrites in the constructions are independent given their context, which is a special case of the sampling assumptions in Section 3.4. Common shifts of the scores within a context leave the summaries unchanged.

Suppose first that $a > 0$. Put $h_0 = \Phi(a,d)$ and define

$$v_- = h_0 - a - d, \qquad v_+ = a - \frac{a^2}{2h_0}. \tag{26}$$

Writing $\delta = \sqrt{d/h_0}$ gives $a = h_0(1-\delta)$, $v_- = h_0\delta(1-\delta)$, and $v_+ = h_0(1-\delta^2)/2$. In particular, $0 \leq v_- \leq v_+ < a$. The first branch increases in v and meets h_0 at v_-; the third decreases and meets h_0 at v_+. Hence

$$\Psi(a,d,v) = \begin{cases} a + d + v, & 0 \leq v \leq v_-, \\ h_0, & v_- \leq v \leq v_+, \\ \big(\sqrt{2(a-v)} + \sqrt{d}\big)^2, & v_+ \leq v < a. \end{cases} \tag{27}$$

First regime: attainment. Let $b = d + v$, $h = a + b$, $w = h$, $\ell = 0$, and $B = 0$. When $ab > 0$, set

$$\lambda = \frac{vh}{ab}, \qquad \Pr(X = h) = \frac{b}{h}, \quad \Pr(X = 0) = \frac{a}{h}, \qquad A = (1-\lambda)b + \lambda X. \tag{28}$$

The condition $v \leq v_-$ is equivalent to $v^2 + dv \leq ad$, so $\lambda \in [0,1]$. Thus $0 \leq A \leq h$, $\mathbb{E}_0 A = b$, and $\frac{1}{2}\mathbb{E}_0|A - A'| = \lambda ab/h = v$. The summaries equal (a,d,v) and $H = h = a + d + v$. If $b = 0$, take $A = B = 0$ and $w = a$.

Second regime: attainment. Take $w = h_0$, $\ell = 0$, and independent $A, B \in \{0, h_0\}$ with $\Pr(A = h_0) = p$, $\Pr(B = h_0) = q$, and $p - q = \delta$. Then

$$a = h_0(1 - \delta), \qquad d = h_0\delta^2, \qquad \bar{\nu}_0 = h_0\{p(1-p) + q(1-q)\}. \tag{29}$$

As q varies over $[0, 1 - \delta]$, the last expression takes every value in $[v_-, v_+]$. Thus every triple in the second regime admits a law with $H = h_0$.

Third regime: an extremizing sequence. Let $v_+ < v < a$ and $e = a - v > 0$. For $t > 0$, define

$$h(t) = \left(\sqrt{2t} + \sqrt{d}\right)^2, \qquad a_0(t) = 2t + \sqrt{2td}. \tag{30}$$

Use endpoints $h(t), 0$ and independent two-point laws satisfying $\Pr(A = 0) = \Pr(B = h(t)) = \sqrt{t/(2h(t))}$. Their summaries are $(a_0(t), d, a_0(t) - t)$ and their gap is $h(t)$. The third-regime condition gives $a_0(e) < a$.

For sufficiently small $\varepsilon > 0$, choose $t_\varepsilon \in (0, e)$ solving

$$t_\varepsilon + \varepsilon\{a - a_0(t_\varepsilon)\} = e. \tag{31}$$

Such a solution exists by continuity: the left side is $\varepsilon a < e$ at $t = 0$ and exceeds e at $t = e$. Moreover, $t_\varepsilon \to e$ as $\varepsilon \downarrow 0$. Put $k_\varepsilon = a - a_0(t_\varepsilon) > 0$. In a second context, let both endpoints be zero and draw A, B independently from a common law placing mass ε at $T_\varepsilon = k_\varepsilon/(2\varepsilon)$ and the remaining mass at zero. This construction has summaries $(k_\varepsilon, 0, k_\varepsilon(1 - \varepsilon))$ and zero gap.

Combine the two constructions by assigning each context probability $1/2$ and doubling its scores. Homogeneity of absolute distances gives aggregate summaries

$$\big(a_0(t_\varepsilon) + k_\varepsilon,\ d,\ a_0(t_\varepsilon) - t_\varepsilon + k_\varepsilon(1 - \varepsilon)\big) = (a, d, v), \tag{32}$$

where the last equality follows from Eq. (31). Every distribution has finite first moments, and

$$H = h(t_\varepsilon) \longrightarrow \left(\sqrt{2(a - v)} + \sqrt{d}\right)^2. \tag{33}$$

Thus the third branch is the supremum, although it need not be attained. For $a = v = 0$, deterministic rewrite scores equal to their respective endpoints, separated by d, give $H = d$. This proves Eq. (22). Finally, $\Psi(a, d, v) < a + d + v$ holds exactly when $v > v_-$, equivalently $d(a - v) < v^2$, establishing Eq. (34).

For $a > \bar{\nu}_0 > 0$, the refinement improves on the cross-path triangle bound exactly when

$$\Psi(a, d, \bar{\nu}_0) < a + d + \bar{\nu}_0 \quad \Longleftrightarrow \quad d(a - \bar{\nu}_0) < \bar{\nu}_0^2. \tag{34}$$

The original-path bound $a + \mathbb{E}_0|g_2|$ remains available and can be combined with Ψ by taking their minimum. Optimality in Eq. (22) does not fix $\mathbb{E}_0|g_2|$, or the separate semantic, length, and twin-noise contributions.

A.5 Remarks on the bound

Since G_θ depends only on Z, its expectation is unchanged by resampling the rewrites. The right-hand quantities are defined under $\mathbb{E}_0$; filtered benchmark averages require a separate justification.

The first term of Eq. (11) is Eq. (6) under the auxiliary law. The second replaces the raw syntactic gap by d, which is the expected Cramér distance between the conditional score laws of syntax-aligned and length-aligned rewrites (Section A.1): the syntactic discrepancy in excess of what twin variability alone produces. Under the sampling assumptions the syntax twins are exchangeable, so $d \geq 0$ states that $\mathbb{E}_0|A_1 - B_2| \geq \bar{\nu}_0$, the cross-twin counterpart of the regularity $S \geq 1$ observed in Section 5.4. When $\bar{\nu}_0$ is fixed as well, Theorem 2 gives the exact worst case and characterizes when it improves on the cross-path triangle bound. The theorem concerns the magnitude of the alignment gap, not its sign, which the directional outcomes of Definition 4 measure. Its sampling law idealizes the construction of Section 4.1, whose two versions of a rewrite are generated together and filtered jointly; on the benchmark we therefore report the empirical analogues of a and d only as diagnostics, a negative estimate of d signalling that the assumptions fail.

A.6 Empirical diagnostics

On the benchmark, with raw log-probabilities and the mechanical term of g_3 retained, $\hat{d} = \frac{1}{2}\mathbb{E}[|A_1 - B_2| + |A_2 - B_2|] - \bar{\nu}$ is positive with a 95% interval excluding zero for all 25 models on the full

set (0.4–0.7 $\bar{\nu}$; Table 3), and for all 22 models with ordinary displacement on both splits and on the 144-family screened subset, and the cross-twin sensitivity $S_\times = \mathbb{E}|A_1 - B_2|/\bar{\nu}$ exceeds 1 for every model, with lower confidence bound at least 1.41. This is the empirical counterpart of $d \geq 0$ and $S \geq 1$: the syntactic discrepancy exceeds twin variability even though the construction violates the i.i.d.-twin assumption, and it does so with independently drawn twins as well (Section F.4). The refined bounds do not bind here: for every model the triangle bound $\sum_j \mathbb{E}|g_j|$ is the smallest of the three, the active branch of Ψ is $a + d + \bar{\nu}$, 0.2–1.1 raw log-probability above it, Φ is about 5 looser, and the improvement condition $d(a - \bar{\nu}) < \bar{\nu}^2$ of Eq. (34) fails by a factor 1.9–3.3 (1.0–2.1 on the screened subset). $\mathbb{E}|G_\theta|$ is 0.44–0.57 of the triangle bound: about half of the summed component magnitudes cancel within families, consistent with Section 5.2. The same holds on the second backbone (13 of 13 models with ordinary displacement, tightness 0.47–0.55).

Table 3: Empirical diagnostics for Theorem 1 on the benchmark (raw log-probability; g_3 with its mechanical term). $\hat{d} = \frac{1}{2}\mathbb{E}[|A_1 - B_2| + |A_2 - B_2|] - \bar{\nu}$ with 95% bootstrap CI; $S_\times = \mathbb{E}|A_1 - B_2|/\bar{\nu}$; tightness $= \mathbb{E}|G_\theta|/\min\{\text{triangle}, \Phi, \Psi\}$. For every model the triangle bound is the smallest of the three, and the improvement condition $\hat{d}(a - \bar{\nu}) < \bar{\nu}^2$ of Eq. (34) fails. †Large displacement.

Model	$\hat{d}$ [CI]	$\hat{d}/\bar{\nu}$	S	$S_\times$ [CI]	tightness [CI]
BASE	4.81 [4.28, 5.34]	0.55	1.521	1.579 [1.508, 1.653]	0.563 [0.542, 0.584]
SOTA methods (92,858 pairs)					
DPO	5.68 [5.03, 6.34]	0.53	1.508	1.545 [1.474, 1.618]	0.544 [0.524, 0.564]
Robust-DPO	5.50 [4.90, 6.10]	0.55	1.520	1.567 [1.496, 1.641]	0.541 [0.520, 0.561]
Dr. DPO	7.31 [6.52, 8.12]	0.52	1.498	1.525 [1.460, 1.595]	0.539 [0.519, 0.559]
R-DPO	5.27 [4.68, 5.85]	0.57	1.539	1.584 [1.511, 1.658]	0.549 [0.528, 0.569]
LD-DPO	5.14 [4.59, 5.71]	0.54	1.506	1.565 [1.494, 1.636]	0.540 [0.520, 0.561]
SamPO	5.07 [4.53, 5.64]	0.54	1.501	1.565 [1.494, 1.636]	0.543 [0.522, 0.563]
SimPO	4.09 [3.64, 4.55]	0.59	1.546	1.618 [1.542, 1.694]	0.565 [0.544, 0.586]
ORPO	4.38 [3.87, 4.91]	0.53	1.495	1.558 [1.489, 1.632]	0.561 [0.539, 0.582]
Tie training	4.75 [4.22, 5.27]	0.53	1.491	1.558 [1.489, 1.626]	0.541 [0.520, 0.561]
LMPO†	4.79 [4.34, 5.22]	0.63	1.577	1.692 [1.622, 1.759]	0.539 [0.521, 0.557]
Dr. DPO variants (same pairs)					
Dr. DPO, β'=0.25†	28.54 [25.55, 31.57]	0.49	1.413	1.552 [1.490, 1.615]	0.449 [0.433, 0.467]
Dr. DPO, β'=0.5†	18.36 [16.40, 20.22]	0.49	1.441	1.537 [1.474, 1.598]	0.496 [0.479, 0.514]
Dr. DPO, β'=2	5.55 [4.93, 6.19]	0.53	1.510	1.547 [1.477, 1.620]	0.542 [0.521, 0.561]
Dr. DPO, length-strat.	6.69 [5.94, 7.48]	0.50	1.479	1.511 [1.448, 1.579]	0.535 [0.516, 0.555]
Dr. DPO, buffered	7.64 [6.81, 8.45]	0.52	1.493	1.521 [1.456, 1.589]	0.538 [0.519, 0.557]
Dr. DPO + R-DPO reg.	8.01 [7.15, 8.84]	0.55	1.520	1.551 [1.485, 1.622]	0.538 [0.519, 0.557]
Dr. DPO + I4	5.39 [4.80, 5.96]	0.66	1.630	1.683 [1.600, 1.763]	0.557 [0.536, 0.578]
Dr. DPO + I4b	5.85 [5.24, 6.45]	0.64	1.618	1.665 [1.588, 1.745]	0.551 [0.530, 0.571]
Counterfactual-data controls (1,998 prompts, 3 seeds)					
Dr. DPO, original pairs	4.60 [4.08, 5.10]	0.54	1.499	1.562 [1.493, 1.635]	0.557 [0.536, 0.578]
DPO, KL-DRO views	4.59 [4.06, 5.09]	0.52	1.485	1.554 [1.485, 1.624]	0.553 [0.531, 0.574]
Dr. DPO, KL-DRO views	4.55 [4.02, 5.05]	0.52	1.479	1.549 [1.480, 1.620]	0.550 [0.529, 0.571]
Dr. DPO, uniform views	4.41 [3.85, 4.95]	0.43	1.393	1.476 [1.415, 1.539]	0.517 [0.497, 0.537]
flattened, 3× steps	8.12 [7.14, 9.08]	0.41	1.357	1.476 [1.417, 1.536]	0.440 [0.422, 0.458]
flattened, step-matched	6.36 [5.55, 7.14]	0.41	1.353	1.476 [1.418, 1.537]	0.461 [0.442, 0.479]

B BENCHMARK CONSTRUCTION AND VALIDATION

B.1 SOURCE DATA AND GENERATION PIPELINE

Source pairs are drawn from the TL;DR human-preference comparisons (Völske et al., 2017; Stiennon et al., 2020). Each source record carries a prompt (a Reddit post in the form `SUBREDDIT / TITLE / POST`), a chosen and a rejected human summary, and their token counts. We sample 2,000 records. Rewrites are produced by an LLM agent pipeline (Claude Code workflow agents, model `claude-opus-5`, default decoding with no temperature or length settings) in shards of 16 families. One *writer* agent per shard writes all four rewrites of every family in one call under the specification below, and one *adversarial auditor* agent per shard checks and repairs all four together before assembly; the released `workflow_scripts/` contain the three scripts used (an initial evaluation build, an initial training build and a resume script that wrote 96 of the 125 evaluation shards and repaired six) with the writer and auditor instructions verbatim. The specification requires every variant to keep the meaning of the rejected summary and never drift toward the chosen one, to be

fluent, to start with one leading space, to have exactly the chosen summary's token count under the Qwen2.5 tokenizer, and to be distinct; it defines y_1'' as "rejected's meaning at N tokens, keeping rejected's own syntactic structure", y_1' as "rejected's meaning at N tokens, mirroring chosen's syntax", instructs that the two "differ only in syntax: reuse the same content words and rearrange them", and asks for an "independent second version" of each. Because the second versions are written in the same call, in a fixed order, they are twins in specification, not independent draws. Token counts are verified with a batched call to the tokenizer; the auditor re-verifies counts, meaning preservation, leakage toward the chosen summary, the syntax contrast and distinctness, and rewrites failures in place.

Two-step construction. The first step aligns length. An LLM writer rewrites the rejected summary so that it has *exactly* the token count of the chosen summary, says the same thing, and keeps the rejected summary's own sentence pattern; only the length changes. The writer produces two versions of this rewrite, y_1'' and y_2'', which share meaning, sentence pattern and token count and differ only in the words chosen. The second step aligns syntax: y_1' rearranges the words of y_1'' into the sentence pattern of the chosen summary at fixed token count, a structural rearrangement rather than a fresh paraphrase; only the syntax changes. The two steps describe the specification; the writer produces all four rewrites of a family at once under it. This again yields two versions, y_1' and y_2'. The path $y_w \to y_1' \to y_1'' \to y_l$ of Eq. (4) thus changes one attribute per step, both intermediate responses have twins, and each intermediate response is a length- or syntax-aligned rewrite in the sense of Definition 1.

B.2 Hard constraints and filtering

The assembler rejects a family if any of the following fails: (T) every rewrite has exactly the chosen summary's token count under the evaluation tokenizer, with zero tolerance, counted on the response normalized to a single leading space; (S) leading whitespace matches the source; (D) no rewrite is byte-identical to the rejected summary or to another rewrite. It also reports, without filtering on it, the character overlap (V) of y_1' and y_1'', which the specification and the audit enforce so that the second step is a structural rearrangement rather than a second paraphrase. Of 2,000 source records, 1,979 pass; all have distinct prompts. Token exactness was re-measured at scoring time: the pairs (chosen, y_1'), (y_1', y_1''), (y_1', y_2') and (y_1'', y_2'') have identical token counts in 100.00% of families, whereas y_1'' and the rejected summary have identical counts in only 4.30% (means 37.1 vs. 35.5 tokens), which is the length channel by construction. Consequently the mechanical term of Eq. (8) is exactly zero for g_1, g_2 and ν, and only g_3 has distinct raw and content versions. The two versions of each rewrite are produced in the same writer call in a fixed order and revised together by the auditor; they are twins in specification, not independent draws (Sections 3.3 and 5.4).

B.3 Validation of the rewrites

LLM-judge screening. 200 families were screened against four rules, each family judged once by one of eight LLM annotators, with no human adjudication and therefore no inter-annotator agreement: y_1' preserves the meaning of the rejected summary; y_1'' preserves it; y_1' does not drift toward the chosen summary; y_1' and y_1'' differ only syntactically. 144 families pass all four; we call this the *screened subset* (it is not human-validated). Table 4 reports the metrics on this subset next to the full set. Absolute values shift (lower M and S, slightly higher L for most models) but the qualitative picture is unchanged: S stays above 1 for every model (1.19–1.47), the I4 variants keep the largest L, and SimPO/ORPO keep L at or below the untrained value. With 144 families the subset is too small to resolve differences between methods: rank correlations between the full set and the subset are 0.81 for L, 0.68 for accuracy, 0.48 for M and 0.12 for S; the last is expected, since S has almost no between-model variance to rank. A human-adjudicated subset of about a thousand families is the first item on our list.

LLM-judge robustness. We also scored every family under three judge conditions on the same four checks, a Claude model under two rubric prompts and Qwen2.5-72B-Instruct (Table 5). Agreement is low: Cohen's κ on "all four checks pass" is 0.155–0.204 between judge families and 0.482 between the two rubrics of the same judge. What follows is therefore a check of stability, not of validity: it shows that the conclusions do not depend on which families a judge would exclude, not that meaning preservation and syntactic contrast are measured accurately, which only blinded

Table 4: Metrics on the full evaluation set (n=1,979) and on the LLM-judge-screened subset (n=144) for representative models.

	full				screened subset			
Model	acc	M	S	L	acc	M	S	L
BASE	0.518	3.238	1.521	1.490	0.549	2.829	1.307	1.650
DPO	0.580	3.535	1.508	1.551	0.618	3.204	1.235	1.613
Robust-DPO	0.568	3.531	1.520	1.536	0.576	3.095	1.191	1.585
Dr. DPO	0.592	3.457	1.498	1.538	0.597	3.318	1.280	1.651
R-DPO	0.565	3.532	1.539	1.661	0.556	2.979	1.227	1.627
LD-DPO	0.548	3.495	1.506	1.492	0.562	3.009	1.218	1.520
SamPO	0.547	3.456	1.501	1.469	0.569	3.008	1.236	1.525
SimPO	0.511	3.329	1.546	1.362	0.549	3.190	1.269	1.413
ORPO	0.514	3.331	1.495	1.486	0.549	2.892	1.211	1.422
Tie training	0.543	3.426	1.491	1.486	0.569	2.990	1.227	1.560
LMPO†	0.531	3.638	1.577	1.208	0.542	3.359	1.466	1.072
Dr. DPO + I4	0.581	3.434	1.630	2.618	0.549	2.949	1.424	2.535
Dr. DPO + I4b	0.589	3.484	1.618	2.330	0.556	3.054	1.374	2.255

human adjudication of a subset can establish. We therefore do not define a single "clean" subset but form six filtered subsets, each condition alone and each pairwise intersection (the pre-registered definition coincides with the intersection of the second Claude rubric and Qwen), retaining between 45.7% and 86.6% of the 1,979 families, and re-run the evaluation on each. Under every definition S exceeds 1 for every model, with the smallest bootstrap lower bound 1.135 on the strictest subset (904 families). Filtering lowers the median S from 1.50 to 1.29–1.45 and widens its median confidence interval from 0.143 to 0.147–0.189. Rankings by the margin-violation rate, by C and by L are stable (Spearman $\rho \geq 0.97$, 0.86 and 0.84 with the full set), whereas the S ranking is not ($\rho = 0.74$–0.98; Table 6), since S has little between-model variance. On the strictest and on the pre-registered subset the flattened-arm gains remain significant and no length-handling objective exceeds DPO, whereas DPO − BASE on C loses significance ($+0.008$ $[-0.013, +0.031]$ on the strictest), as does, on the swapped path, the R-DPO advantage of Table 12.

Table 5: LLM-judge screening of all 1,979 families. Top: pass rate of each of the four checks and of "all four" per judge condition, with the retained family count. Bottom: Cohen's κ between judge conditions (last column: raw agreement on "all four").

Judge	check 1	check 2	check 3	check 4	all four	n
Claude, rubric 1	0.792	0.928	0.874	0.808	0.725	1,434
Claude, rubric 2	0.961	0.988	0.953	0.883	0.866	1,714
Qwen2.5-72B	0.809	0.832	0.787	0.728	0.564	1,116
κ	check 1	check 2	check 3	check 4	all four	agree
Claude 1 vs Claude 2	0.254	0.250	0.421	0.575	0.482	0.826
Claude 1 vs Qwen	0.191	0.159	0.094	0.165	0.204	0.625
Claude 2 vs Qwen	0.114	0.058	0.072	0.155	0.155	0.617

B.4 SPLITS AND UNCERTAINTY

Families are assigned to development and test splits deterministically by `md5(prompt) % 100 < 40` (dev: 802 families; test: 1,177; full: 1,979). Any selection (model choice, thresholds, figure design) is made on dev only; the main text reports the full set and labels it as such; every contrast and verdict in this paper is unchanged on the test split (released with the score files). Confidence intervals are 95% percentile intervals from a non-parametric bootstrap over families (2,000 resamples, seed 0, resampling with replacement to the original size); all metrics of a model are computed on the same resample so that differences between metrics inherit paired variability. Configurations with

Table 6: Spearman rank correlation between the 22 models' ranking on the full set and on each judge-filtered subset (models with large displacement excluded from the ranking); 95% bootstrap CI over families for S. Rankings by the margin-violation rate $V = \Pr(|g_1| \le |g_2| + |g_3|)$, by C and by L are stable; the S ranking is not, because S has little between-model variance. *Coincides with the pre-registered definition.

Subset	n	S [CI]	V	C	L	M/S
Claude, rubric 1	1,434	0.94 [0.78, 0.98]	0.996	0.958	0.855	0.842
Claude, rubric 2	1,714	0.97 [0.88, 0.99]	0.992	0.984	0.980	0.983
Qwen2.5-72B	1,116	0.77 [0.55, 0.94]	0.987	0.942	0.982	0.894
Claude 1 ∩ Claude 2	1,402	0.93 [0.77, 0.98]	0.997	0.962	0.848	0.829
Claude 1 ∩ Qwen	904	0.73 [0.36, 0.91]	0.981	0.860	0.939	0.732
Claude 2 ∩ Qwen*	1,036	0.76 [0.49, 0.93]	0.976	0.959	0.976	0.810

several seeds are reported as seed means; the seed spread measures training randomness and is not a confidence interval.

C Models, training and evaluation protocol

C.1 Backbone and scoring

All models are Qwen2.5-1.5B-Instruct (Yang et al., 2024) or fine-tunes of it; BASE is the untrained instruct model, the reference π_{ref} and the initialization of every trained model. A response is scored by the sum of token log-probabilities of the response conditioned on the prompt under the model's own tokenizer; token counts used in Eq. (8) are recomputed at scoring time on the exact scored string. The second-backbone study (Section F.5) uses Llama-3.2-3B-Instruct in the same role.

C.2 SOTA methods (10 models)

One model per method, trained on the 92,858-pair TL;DR preference training split (Stiennon et al., 2020): DPO, Robust-DPO, Dr. DPO, R-DPO, LD-DPO, SamPO, SimPO, ORPO, tie training and LMPO, with the configurations of Table 7. DPO uses $\beta = 0.05$ whereas the other DPO-family objectives use $\beta = 0.1$. The sweep of Table 14 suggests that this leaves C unchanged (LD-DPO at $\alpha = 0.8$, close to DPO with $\beta = 0.1$, has $C = 0.794$ against DPO's 0.790) but lowers acc and L, so the accuracy differences between DPO and the length-handling objectives in Table 2 may partly reflect β; a DPO run at $\beta = 0.1$ would settle this directly and is not part of this version.

C.3 Dr.DPO variants (8 models)

Same training data as above; definitions read from the training code (Table 7). Dr. DPO aggregates the per-pair DPO losses ℓ_i of a micro-batch of eight by the KL-DRO dual $-\beta' \log \frac{1}{8} \sum_i \exp(-\ell_i/\beta')$, whose gradient weights the pairs by $\mathrm{softmax}_i(-\ell_i/\beta')$: low-loss pairs are up-weighted and high-loss pairs down-weighted, more sharply for smaller β'. The variants change $\beta' \in \{0.25, 0.5, 2\}$; compute the dual separately within the chosen-longer and the other pairs of each micro-batch and average the two (length-stratified); normalize the weights over the micro-batch plus a first-in-first-out buffer of the last 128 losses (buffered); insert R-DPO's term $-\alpha(l(y_w) - l(y_l))$ with $\alpha = 0.05$ inside each pair's logit before the dual; or add λ times the squared difference between the mean raw policy margin of the chosen-longer and of the other pairs in the micro-batch, with $\lambda = 5 \cdot 10^{-4}$ (I4) and $2.5 \cdot 10^{-4}$ (I4b).

C.4 Counterfactual-data controls (6 arms × 3 seeds)

Trained on the separate 1,998-prompt training set whose prompts are disjoint from the evaluation prompts (verified by exact prompt-string comparison; the `source_id` field is not a global key and is not used for joins), with $\beta = 0.1$, three seeds each, reported as seed means (Table 7). Every

Table 7: Training configurations of the 25 Qwen2.5-1.5B-Instruct models (37 runs), read from each checkpoint's saved training arguments. Common to all: full fine-tuning, bf16, learning rate $5 \cdot 10^{-6}$ with cosine decay and 10% warm-up, per-device batch 8, max length 1,024, one GPU per run. β is the DPO-family temperature (SimPO and LMPO: their own β). Tie training runs one epoch over a $2\times$ mix of strict and tied pairs. Wall-clock in hours per run (controls: mean over seeds).

Model	Objective as implemented	β	epochs	steps	eff. batch	seed(s)	h
SOTA methods (92,858 pairs)							
DPO	TRL `sigmoid` loss	0.05	2	11,608	16	42	4.3
Robust-DPO	TRL `robust` loss, label smoothing 0.1	0.1	2	11,608	16	42	2.0
Dr. DPO	KL-DRO dual over the micro-batch, $\beta' = 1$	0.1	2	11,608	16	42	1.8
R-DPO	$-\alpha(l(y_w) - l(y_l))$ in the logit, $\alpha = 0.05$	0.1	2	11,608	16	42	1.8
LD-DPO	excess-length tokens down-weighted, $\alpha = 0.5$	0.1	2	11,608	16	42	2.0
SamPO	token log-ratios down-sampled to $\min(l(y_w), l(y_l))$	0.1	2	11,608	16	42	1.8
SimPO	TRL CPO/SimPO, $\gamma = 1.4$, reference-free	2.5	2	11,608	16	42	3.5
ORPO	TRL ORPO, $\lambda = 0.1$	0.1	2	11,608	16	42	1.5
Tie training	tie weight 1, 5-token tolerance, $2\times$ data	0.1	1	11,608	16	42	1.8
LMPO	$\lambda = 0.2$, $k = 5$, $\gamma = 0.55\beta$, reference-free	2.5	2	11,608	16	42	1.3
Dr. DPO variants (92,858 pairs)							
Dr. DPO, β'=0.25	sharper DRO reweighting	0.1	2	11,608	16	42	2.7
Dr. DPO, β'=0.5	sharper DRO reweighting	0.1	2	11,608	16	42	1.8
Dr. DPO, β'=2	flatter reweighting	0.1	2	11,608	16	42	1.9
Dr. DPO, length-stratified	dual within chosen-longer / other pairs	0.1	2	11,608	16	42	2.7
Dr. DPO, buffered	weights normalised over a 128-loss buffer	0.1	2	11,608	16	42	1.8
Dr. DPO + R-DPO reg.	R-DPO term in each pair loss, $\alpha = 0.05$	0.1	2	11,608	16	42	1.9
Dr. DPO + I4	group-mean length-bias penalty, $\lambda = 5 \cdot 10^{-4}$	0.1	2	11,608	16	42	1.8
Dr. DPO + I4b	same, $\lambda = 2.5 \cdot 10^{-4}$	0.1	2	11,608	16	42	1.9
Counterfactual-data controls (1,998 prompts)							
Dr. DPO, original pairs	$\beta' = 1$, 1,998 pairs	0.1	2	250	16	1, 2, 3	0.1
DPO, KL-DRO over views	inner KL-DRO, $p = (0.85, 0.10, 0.05)$, $\tau = 1$	0.1	2	250	16	1, 2, 3	0.1
Dr. DPO, KL-DRO over views	same, outer Dr. DPO dual	0.1	2	250	16	1, 2, 3	0.1
Dr. DPO, uniform views	nominal $p = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, $\tau = 1$	0.1	2	250	16	1, 2, 3	0.1
flattened pairs, $3\times$ steps	Dr. DPO $\beta' = 1$ on 5,994 pairs	0.1	2	750	16	1, 2, 3	0.2
flattened pairs, step-matched	same, gradient accumulation 6	0.1	2	250	48	1, 2, 3	0.1

training example provides three views that share the chosen summary: the original pair, the length-matched pair (y_w, y_1'') and the length- and syntax-matched pair (y_w, y_1'). *Dr. DPO, original pairs* uses the original pair only (250 steps). The three *KL-DRO* arms combine the three per-view losses of each example by an inner aggregate $\tau \log \sum_k p_k \exp(\ell_k/\tau)$ with $\tau = 1$, which tilts the nominal view weights p toward the highest-loss view, and then average the examples (DPO) or apply the Dr. DPO dual over them (Dr. DPO); the nominal weights are $(0.85, 0.10, 0.05)$ for the original, length-matched and syntax-matched view, or uniform $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ for the *uniform views* arm (250 steps each). The two *flattened* arms train Dr. DPO on the 5,994 views as independent pairs: *step-matched* uses gradient accumulation 6 (effective batch 48) for 250 steps, and $3\times$ *steps* the standard batch for 750 steps. The controls see 46 times fewer training pairs than the other groups, so only within-group comparisons are attributed.

C.5 Large likelihood displacement

A model shows large displacement when the average log-probability of the chosen summaries has fallen by more than 60 relative to BASE ($\Delta_{\text{disp}} < -60$). Three Qwen models trigger the flag, LMPO and Dr. DPO with $\beta' = 0.25$ and 0.5, and three Llama models, SimPO, LMPO and the $3\times$-step flattened arm. Their pairwise accuracy is ordinary (Dr. DPO with $\beta' = 0.5$ has the highest accuracy of all Qwen models), so we treat them as regular training outcomes and include them in every table and count, marking them with †. What the flag changes is the ruler: the two sharp-DRO variants have twin scales of 37 and 58 raw log-probability against 8.7 for BASE, so their S and L in units of ν are compressed while M is not, and the three flagged Qwen models are the only ones whose semantic gap couples more with the length than with the syntactic residual (Fig. 3).

D DIRECTIONAL SEMANTIC OUTCOME AND PER-SAMPLE ANALYSIS

Provenance. The analyses of Sections 5.3 and 5.4 use the 37 per-family score files (one per model and seed) from which Table 2 was computed, dated before the table was frozen; recomputing accuracy, M, S, L and Δ_{disp} from these files reproduces every entry of Table 2 to within 10^{-14}, and re-tokenizing the 1,979 $\times$ 6 scored strings reproduces the stored token counts exactly. A SHA256 manifest of the 37 files and the MD5 of the frozen table are released with the data.

Definitions. The unsigned separation M does not distinguish a model that prefers the chosen meaning from one that prefers the rejected meaning. We therefore report, for every model, the directional correctness $C = \Pr(g_1 > 0)$, the signed semantic gap $\bar{g}_1 = \mathbb{E}\, g_1/\bar{\nu}$ and the logistic loss $\mathbb{E}[-\log \sigma(g_1)]$, each with 95% bootstrap intervals over families (2,000 resamples, seed 0), together with paired same-family differences against DPO and against BASE. The signed gap accounts for 70–85% of the unsigned separation across models ($\bar{g}_1/M = 0.70$–0.85), so most of M is directional signal rather than spread; the unsigned quantity is nevertheless not a measure of alignment. For each model we also compute, over its 1,979 families, the Pearson and Spearman correlations of g_1 with g_2 and with g_3^c, the partial correlations, the standardized regression of g_1 on both, and the bootstrap interval of $\text{corr}(g_1, g_2) - \text{corr}(g_1, g_3^c)$, repeated with 5% winsorized correlations and on the screened subset. The script `analysis/reanalysis.py` computes all of these from the per-family score files.

Results. Table 8 gives the levels for BASE and DPO, the DPO–BASE contrast, the seven length-handling objectives paired against DPO and the counterfactual-control contrasts; Table 9 lists every model. Relative to DPO, no length-handling objective improves C or $\bar{g}_1$; SimPO, ORPO, tie training and SamPO are significantly lower on $\bar{g}_1$ and SimPO also on C. Relative to BASE, SimPO and ORPO do not improve C (-0.003 and $+0.011$, both intervals containing zero), whereas every DPO-family objective does ($+0.015$ to $+0.023$). Under the designated first-rewrite, length-first path, the contrast $\text{corr}(g_1, g_2) - \text{corr}(g_1, g_3^c)$ is positive with an interval excluding zero in 17 of 25 models and inconclusive in five; the three significantly negative contrasts are the large-displacement models. Regressions of g_1 on g_2 and g_3^c reach $R^2 = 0.001$–0.112. BASE has $r(g_1, g_2) = 0.286$ and a contrast $\Delta r = +0.234$ $[+0.162, +0.307]$; relative to BASE, 20 of 24 trained models have significantly lower $r(g_1, g_2)$, four are indistinguishable and none is higher. Spearman contrasts are positive in 13 and negative in 12 models; 5% winsorized Pearson contrasts are positive in 18 and negative in 7; on the screened subset no model is significantly positive, two (large displacement) are significantly negative and the other 22 are inconclusive. Tables 11 and 12 give every model's C and S, and every contrast, under the swapped and the symmetrized twin designation (Section 5.4).

Absolute magnitudes. Table 10 gives the twin-discrepancy scale and the mean absolute gaps in raw log-probability, and S and L on a common scale, the untrained model's $\bar{\nu}$. The absolute magnitudes move together with the scale: across the 25 models $\bar{\nu}$ ranges from 6.9 (SimPO) to 58.4 (Dr. DPO, $\beta' = 0.25$) and $\mathbb{E}|g_2|$ from 10.7 to 82.5, with Pearson $r = 0.999$ between the two (0.98 with $\mathbb{E}|g_3^c|$), so on the common scale the models are ranked by their overall score scale rather than by anything specific to syntax: S_{base} runs from 1.24 to 9.49. Six models (SimPO, ORPO, LMPO and three control arms) have $\mathbb{E}|g_2|$ below the untrained 13.2 while their S is at or near the untrained value, because every score difference shrinks or grows with the model. The invariance of Section 5.4 is therefore a statement about the ratio, the quantity Eq. (9) defines, and the lower S of the flattened controls is a change of ratio, their $\mathbb{E}|g_2|$ being the largest of the six arms; the directional outcomes C and $\bar{g}_1$ do not depend on this choice.

Table 8: Directional outcomes on the 1,979 families: directional correctness $C = \Pr(g_1 > 0)$ and signed semantic gap $\bar{g}_1$. The first two rows are levels; the others are paired same-family differences with 95% bootstrap CIs (2,000 resamples over families; intervals for the seed-averaged controls hold the training seeds fixed). †Large displacement. No length-handling objective improves on DPO; flattening the rewrites into ordinary pairs improves both outcomes.

Comparison	C	$\bar{g}_1$
BASE, level	0.774	2.36
DPO, level	0.790	2.73
DPO − BASE	$+0.016$ $[+0.002, +0.031]$	$+0.367$ $[+0.277, +0.452]$
Length-handling objectives, paired against DPO		
R-DPO − DPO	$+0.003$ $[-0.007, +0.014]$	$+0.012$ $[-0.035, +0.061]$
LD-DPO − DPO	$+0.004$ $[-0.006, +0.014]$	-0.041 $[-0.092, +0.013]$
SamPO − DPO	$+0.002$ $[-0.008, +0.013]$	-0.088 $[-0.144, -0.030]$
SimPO − DPO	-0.019 $[-0.031, -0.007]$	-0.290 $[-0.367, -0.214]$
ORPO − DPO	-0.006 $[-0.020, +0.009]$	-0.206 $[-0.294, -0.120]$
Tie training − DPO	-0.001 $[-0.012, +0.011]$	-0.097 $[-0.163, -0.029]$
LMPO† − DPO	-0.036 $[-0.056, -0.016]$	-0.163 $[-0.342, +0.017]$
Counterfactual-data controls (same 1,998 prompts, 3 seeds)		
flattened pairs, 3× steps − Dr. DPO, original pairs	$+0.059$ $[+0.044, +0.073]$	$+0.301$ $[+0.205, +0.396]$
flattened pairs, step-matched − Dr. DPO, original pairs	$+0.045$ $[+0.032, +0.058]$	$+0.215$ $[+0.140, +0.294]$
flattened pairs, 3× steps − Dr. DPO, KL-DRO over views	$+0.053$ $[+0.040, +0.068]$	$+0.272$ $[+0.178, +0.364]$
flattened pairs, step-matched − Dr. DPO, KL-DRO over views	$+0.039$ $[+0.027, +0.052]$	$+0.186$ $[+0.114, +0.261]$

Table 9: Per-model directional outcomes and per-family associations on the 1,979 families (first-rewrite, length-first path). C with 95% bootstrap CI; $\bar{g}_1$; M, S, L; the twin-discrepancy scale $\bar{\nu}$ in raw log-probability; the within-model Pearson correlations $r(g_1, g_2)$ / $r(g_1, g_3^c)$; and their difference Δr with CI. †large displacement. Controls are seed means over three seeds. Paired differences against DPO and BASE for every model are in the released `paired_diffs.csv`.

Model	C [CI]	$\bar{g}_1$	M	S	L	$\bar{\nu}$	$r(g_1, g_2)$ / $r(g_1, g_3^c)$	Δr [CI]
BASE (untrained)	0.774 [0.754, 0.792]	2.36	3.238	1.521	1.490	8.7	0.286 / 0.052	$+0.234$ [$+0.162$, $+0.307$]
SOTA methods (92,858 training pairs)								
DPO	0.790 [0.771, 0.807]	2.73	3.535	1.508	1.551	10.7	0.200 / 0.053	$+0.147$ [$+0.063$, $+0.226$]
Robust-DPO	0.796 [0.778, 0.813]	2.73	3.531	1.520	1.536	10.0	0.207 / 0.048	$+0.159$ [$+0.076$, $+0.238$]
Dr. DPO	0.790 [0.772, 0.807]	2.70	3.457	1.498	1.538	14.1	0.127 / 0.047	$+0.080$ [-0.006, $+0.159$]
R-DPO	0.793 [0.774, 0.810]	2.74	3.532	1.539	1.661	9.3	0.238 / 0.050	$+0.188$ [$+0.110$, $+0.260$]
LD-DPO	0.794 [0.776, 0.811]	2.69	3.495	1.506	1.492	9.5	0.212 / 0.048	$+0.164$ [$+0.087$, $+0.241$]
SamPO	0.792 [0.774, 0.809]	2.64	3.456	1.501	1.469	9.4	0.211 / 0.045	$+0.166$ [$+0.090$, $+0.243$]
SimPO	0.771 [0.751, 0.788]	2.44	3.329	1.546	1.362	6.9	0.263 / 0.036	$+0.227$ [$+0.149$, $+0.304$]
ORPO	0.784 [0.765, 0.802]	2.52	3.331	1.495	1.486	8.3	0.325 / 0.070	$+0.255$ [$+0.187$, $+0.321$]
Tie training	0.789 [0.771, 0.806]	2.63	3.426	1.491	1.486	9.0	0.241 / 0.056	$+0.184$ [$+0.108$, $+0.259$]
LMPO†	0.754 [0.735, 0.772]	2.56	3.638	1.577	1.208	7.6	-0.087 / 0.051	-0.138 [-0.213, -0.063]
Dr. DPO variants (same training pairs)								
Dr. DPO, $\beta'=0.25$†	0.766 [0.746, 0.783]	2.36	3.237	1.413	1.213	58.4	-0.073 / 0.040	-0.114 [-0.180, -0.042]
Dr. DPO, $\beta'=0.5$†	0.774 [0.754, 0.791]	2.39	3.231	1.441	1.290	37.4	-0.068 / 0.108	-0.176 [-0.255, -0.098]
Dr. DPO, $\beta'=2$	0.792 [0.775, 0.810]	2.74	3.552	1.510	1.548	10.4	0.196 / 0.048	$+0.147$ [$+0.065$, $+0.226$]
Dr. DPO, length-stratified	0.797 [0.779, 0.814]	2.73	3.506	1.479	1.554	13.3	0.129 / 0.045	$+0.084$ [-0.002, $+0.167$]
Dr. DPO, buffered	0.794 [0.776, 0.812]	2.66	3.418	1.493	1.544	14.7	0.117 / 0.049	$+0.068$ [-0.020, $+0.150$]
Dr. DPO + R-DPO regularizer	0.793 [0.775, 0.811]	2.70	3.472	1.520	1.654	14.7	0.111 / 0.049	$+0.062$ [-0.024, $+0.140$]
Dr. DPO + I4	0.795 [0.777, 0.813]	2.68	3.434	1.630	2.618	8.2	0.294 / 0.073	$+0.222$ [$+0.139$, $+0.297$]
Dr. DPO + I4b	0.793 [0.775, 0.811]	2.71	3.484	1.618	2.330	9.1	0.256 / 0.072	$+0.185$ [$+0.102$, $+0.259$]
Counterfactual-data controls (1,998 prompts, 3 seeds; seed means)								
Dr. DPO, original pairs	0.775 [0.757, 0.794]	2.42	3.279	1.499	1.479	8.6	0.276 / 0.055	$+0.220$ [$+0.150$, $+0.291$]
DPO, KL-DRO over views	0.781 [0.762, 0.800]	2.43	3.273	1.485	1.489	8.8	0.271 / 0.058	$+0.213$ [$+0.143$, $+0.285$]
Dr. DPO, KL-DRO over views	0.781 [0.763, 0.800]	2.44	3.277	1.479	1.487	8.8	0.267 / 0.060	$+0.207$ [$+0.136$, $+0.278$]
Dr. DPO, uniform views	0.796 [0.777, 0.814]	2.49	3.205	1.393	1.521	10.4	0.205 / 0.061	$+0.144$ [$+0.075$, $+0.214$]
flattened pairs, 3× steps	0.834 [0.818, 0.850]	2.72	3.203	1.357	1.571	19.7	0.025 / -0.027	$+0.052$ [-0.018, $+0.123$]
flattened pairs, step-matched	0.820 [0.804, 0.836]	2.63	3.182	1.353	1.556	15.7	0.075 / -0.006	$+0.081$ [$+0.011$, $+0.149$]

Table 10: Absolute magnitudes behind the scale-normalized metrics: the twin-discrepancy scale $\bar{\nu}$ and the mean absolute gaps $\mathbb{E}|g_1|$, $\mathbb{E}|g_2|$, $\mathbb{E}|g_3^c|$ of Table 2 in raw log-probability, together with S and its counterpart on a common scale, the untrained model's $\bar{\nu}$ ($S_{\text{base}} = \mathbb{E}|g_2|/\bar{\nu}_{\text{BASE}}$, L_{base} likewise). $\bar{\nu}$ grows with training, so the ratios of Table 2 and the absolute magnitudes can move in opposite directions; Section 5.4 and Section 6 discuss which of the two a claim is about. †Large displacement.

Model	$\bar{\nu}$	$\mathbb{E}\|g_1\|$	$\mathbb{E}\|g_2\|$	$\mathbb{E}\|g_3^c\|$	S	S_{base}	L_{base}
BASE	8.7	28.1	13.2	12.9	1.521	1.52	1.49
SOTA methods (92,858 training pairs)							
DPO	10.7	37.8	16.1	16.6	1.508	1.85	1.91
Robust-DPO	10.0	35.4	15.2	15.4	1.520	1.75	1.77
Dr. DPO	14.1	48.6	21.1	21.6	1.498	2.42	2.49
R-DPO	9.3	32.7	14.3	15.4	1.539	1.64	1.77
LD-DPO	9.5	33.2	14.3	14.2	1.506	1.65	1.63
SamPO	9.4	32.4	14.1	13.8	1.501	1.62	1.59
SimPO	6.9	23.1	10.7	9.5	1.546	1.24	1.09
ORPO	8.3	27.7	12.4	12.3	1.495	1.43	1.42
Tie training	9.0	31.0	13.5	13.4	1.491	1.55	1.55
LMPO†	7.6	27.8	12.0	9.2	1.577	1.39	1.06
Dr. DPO variants (same training pairs)							
Dr. DPO, $\beta'=0.25^{\dagger}$	58.4	188.9	82.5	70.8	1.413	9.49	8.15
Dr. DPO, $\beta'=0.5^{\dagger}$	37.4	120.8	53.9	48.2	1.441	6.21	5.55
Dr. DPO, $\beta'=2$	10.4	37.0	15.7	16.1	1.510	1.81	1.86
Dr. DPO, length-stratified	13.3	46.7	19.7	20.7	1.479	2.27	2.39
Dr. DPO, buffered	14.7	50.3	22.0	22.7	1.493	2.53	2.62
Dr. DPO + R-DPO reg.	14.7	51.1	22.4	24.3	1.520	2.57	2.80
Dr. DPO + I4	8.2	28.2	13.4	21.5	1.630	1.54	2.47
Dr. DPO + I4b	9.1	31.8	14.8	21.3	1.618	1.70	2.45
Counterfactual-data controls (1,998 prompts, 3 seeds)							
Dr. DPO, original pairs	8.6	28.1	12.9	12.7	1.499	1.48	1.46
DPO, KL-DRO over views	8.8	28.7	13.0	13.1	1.485	1.50	1.50
Dr. DPO, KL-DRO over views	8.8	28.9	13.0	13.1	1.479	1.50	1.51
Dr. DPO, uniform views	10.4	33.2	14.4	15.8	1.393	1.66	1.81
flattened pairs, 3× steps	19.7	63.1	26.7	30.9	1.357	3.08	3.56
flattened pairs, step-matched	15.7	50.0	21.2	24.4	1.353	2.44	2.81

Table 11: Directional correctness C and syntactic sensitivity S under three twin designations (Section 5.4): the designated path (y_1', y_1''), the swapped path (y_2', y_2'') and the symmetrized path (twin-mean scores). Δ: paired same-family difference swapped $-$ designated with 95% bootstrap CI (2,000 resamples). †Large displacement.

	C				S			
Model	desig.	swapped	symm.	Δ [CI]	desig.	swapped	symm.	Δ [CI]
BASE	0.774	0.816	0.804	+0.042 [+0.029, +0.056]	1.521	1.528	1.446	+0.007 [−0.039, +0.052]
SOTA methods (92,858 pairs)								
DPO	0.790	0.830	0.827	+0.040 [+0.028, +0.054]	1.508	1.518	1.427	+0.010 [−0.039, +0.055]
Robust-DPO	0.796	0.839	0.821	+0.043 [+0.031, +0.056]	1.520	1.531	1.446	+0.011 [−0.039, +0.057]
Dr. DPO	0.790	0.829	0.821	+0.039 [+0.026, +0.052]	1.498	1.515	1.402	+0.017 [−0.033, +0.064]
R-DPO	0.793	0.842	0.822	+0.049 [+0.036, +0.062]	1.539	1.553	1.468	+0.014 [−0.033, +0.060]
LD-DPO	0.794	0.832	0.821	+0.038 [+0.026, +0.051]	1.506	1.518	1.431	+0.011 [−0.037, +0.058]
SamPO	0.792	0.828	0.816	+0.036 [+0.024, +0.050]	1.501	1.517	1.428	+0.015 [−0.032, +0.062]
SimPO	0.771	0.814	0.802	+0.043 [+0.031, +0.057]	1.546	1.560	1.473	+0.014 [−0.035, +0.059]
ORPO	0.784	0.838	0.823	+0.054 [+0.041, +0.067]	1.495	1.497	1.421	+0.002 [−0.040, +0.043]
Tie training	0.789	0.834	0.820	+0.045 [+0.032, +0.059]	1.491	1.493	1.414	+0.002 [−0.044, +0.047]
LMPO†	0.754	0.760	0.759	+0.007 [−0.007, +0.020]	1.577	1.561	1.453	−0.016 [−0.071, +0.038]
Dr. DPO variants (same pairs)								
Dr. DPO, β'=0.25†	0.766	0.811	0.792	+0.045 [+0.031, +0.058]	1.413	1.426	1.313	+0.013 [−0.037, +0.062]
Dr. DPO, β'=0.5†	0.774	0.808	0.798	+0.034 [+0.021, +0.046]	1.441	1.445	1.329	+0.004 [−0.049, +0.056]
Dr. DPO, β'=2	0.792	0.835	0.827	+0.042 [+0.030, +0.055]	1.510	1.520	1.431	+0.011 [−0.040, +0.055]
Dr. DPO, length-strat.	0.797	0.837	0.822	+0.040 [+0.029, +0.052]	1.479	1.493	1.390	+0.015 [−0.036, +0.060]
Dr. DPO, buffered	0.794	0.827	0.820	+0.033 [+0.019, +0.046]	1.493	1.517	1.400	+0.025 [−0.027, +0.073]
Dr. DPO + R-DPO reg.	0.793	0.826	0.812	+0.033 [+0.021, +0.046]	1.520	1.538	1.425	+0.019 [−0.032, +0.067]
Dr. DPO + I4	0.795	0.841	0.829	+0.046 [+0.034, +0.059]	1.630	1.632	1.546	+0.001 [−0.043, +0.049]
Dr. DPO + I4b	0.793	0.843	0.823	+0.050 [+0.037, +0.063]	1.618	1.616	1.528	−0.002 [−0.050, +0.045]
Counterfactual-data controls (1,998 prompts, 3 seeds)								
Dr. DPO, original pairs	0.775	0.823	0.805	+0.048 [+0.034, +0.061]	1.499	1.509	1.427	+0.010 [−0.035, +0.055]
DPO, KL-DRO views	0.781	0.823	0.808	+0.042 [+0.029, +0.056]	1.485	1.493	1.411	+0.008 [−0.038, +0.054]
Dr. DPO, KL-DRO views	0.781	0.825	0.811	+0.044 [+0.031, +0.058]	1.479	1.484	1.405	+0.005 [−0.041, +0.051]
Dr. DPO, uniform views	0.796	0.845	0.826	+0.049 [+0.037, +0.062]	1.393	1.375	1.298	−0.018 [−0.061, +0.027]
flattened, 3× steps	0.834	0.870	0.863	+0.036 [+0.024, +0.047]	1.357	1.349	1.252	−0.008 [−0.051, +0.035]
flattened, step-matched	0.820	0.859	0.847	+0.038 [+0.027, +0.050]	1.353	1.334	1.247	−0.019 [−0.061, +0.023]

Table 12: The paired contrasts of Table 8 under the designated, the swapped and the symmetrized twin designation (95% bootstrap CIs, same families). The significance class changes in 6 of the 24 entries under the swap and in 3 under symmetrization; DPO $>$ BASE, SimPO $<$ DPO and the flattened-arm gains hold under every designation. †Large displacement.

Contrast		designated	swapped	symmetrized
DPO − BASE	C	+0.016 [+0.002, +0.031]	+0.014 [+0.001, +0.028]	+0.022 [+0.009, +0.035]
	$\bar{g}_1$	+0.367 [+0.277, +0.452]	+0.285 [+0.198, +0.374]	+0.326 [+0.237, +0.410]
R-DPO − DPO	C	+0.003 [−0.007, +0.014]	+0.012 [+0.003, +0.021]	−0.005 [−0.014, +0.004]
	$\bar{g}_1$	+0.012 [−0.035, +0.061]	+0.071 [+0.022, +0.120]	+0.041 [−0.006, +0.091]
LD-DPO − DPO	C	+0.004 [−0.006, +0.014]	+0.002 [−0.008, +0.011]	−0.006 [−0.016, +0.004]
	$\bar{g}_1$	−0.041 [−0.092, +0.013]	+0.002 [−0.050, +0.058]	−0.019 [−0.071, +0.034]
SamPO − DPO	C	+0.002 [−0.008, +0.013]	−0.002 [−0.012, +0.008]	−0.011 [−0.022, −0.001]
	$\bar{g}_1$	−0.088 [−0.144, −0.030]	−0.047 [−0.104, +0.012]	−0.068 [−0.124, −0.009]
SimPO − DPO	C	−0.019 [−0.031, −0.007]	−0.016 [−0.028, −0.005]	−0.025 [−0.037, −0.013]
	$\bar{g}_1$	−0.290 [−0.367, −0.214]	−0.264 [−0.344, −0.187]	−0.277 [−0.355, −0.202]
ORPO − DPO	C	−0.006 [−0.020, +0.009]	+0.008 [−0.005, +0.021]	−0.004 [−0.016, +0.010]
	$\bar{g}_1$	−0.206 [−0.294, −0.120]	−0.075 [−0.165, +0.015]	−0.141 [−0.228, −0.055]
Tie training − DPO	C	−0.001 [−0.012, +0.011]	+0.004 [−0.008, +0.015]	−0.007 [−0.019, +0.004]
	$\bar{g}_1$	−0.097 [−0.163, −0.029]	−0.024 [−0.092, +0.045]	−0.061 [−0.127, +0.008]
LMPO† − DPO	C	−0.036 [−0.056, −0.016]	−0.070 [−0.089, −0.051]	−0.067 [−0.086, −0.049]
	$\bar{g}_1$	−0.163 [−0.342, +0.017]	−0.543 [−0.709, −0.367]	−0.353 [−0.523, −0.177]
flattened (3×) − original pairs	C	+0.059 [+0.044, +0.073]	+0.047 [+0.034, +0.060]	+0.057 [+0.044, +0.071]
	$\bar{g}_1$	+0.301 [+0.205, +0.396]	+0.273 [+0.180, +0.368]	+0.287 [+0.193, +0.382]
flattened (step-m.) − original pairs	C	+0.045 [+0.032, +0.058]	+0.036 [+0.025, +0.047]	+0.042 [+0.030, +0.055]
	$\bar{g}_1$	+0.215 [+0.140, +0.294]	+0.200 [+0.127, +0.278]	+0.207 [+0.135, +0.286]
flattened (3×) − Dr. DPO, KL-DRO	C	+0.053 [+0.040, +0.068]	+0.045 [+0.032, +0.058]	+0.052 [+0.039, +0.065]
	$\bar{g}_1$	+0.272 [+0.178, +0.364]	+0.242 [+0.152, +0.334]	+0.257 [+0.166, +0.348]
flattened (step-m.) − Dr. DPO, KL-DRO	C	+0.039 [+0.027, +0.052]	+0.034 [+0.022, +0.045]	+0.037 [+0.025, +0.049]
	$\bar{g}_1$	+0.186 [+0.114, +0.261]	+0.168 [+0.100, +0.243]	+0.177 [+0.109, +0.252]

E CROSS-MODEL VIEW

Figure 3 shows the per-model coupling contrast of Section 5.4, Fig. 4 plots S against L (Section 5.2) and Fig. 5 shows the three metrics as strips. Figure 6 shows the cross-model relation of M to S and L. Across the 25 models M correlates with S (Pearson 0.67) and not with L (0.19), but this reflects training regimes: in raw log-probability the three magnitudes scale together with the model's twin scale (Pearson 1.00 between $\mathbb{E}|g_1|$ and $\mathbb{E}|g_2|$ and 0.98 between $\mathbb{E}|g_1|$ and $\mathbb{E}|g_3^c|$), among the 14 same-data models with ordinary displacement and no mean-length penalty S barely varies ($r(M,S) = -0.12$) while M rises with L ($+0.57$), and across the 23 models other than the I4 variants, whose penalty raises S without moving M, it is $+0.73$ $[+0.47, +0.89]$ (Fig. 2; $+0.70$ without the two sharp-DRO variants, $+0.60$ without the six controls). The per-family analysis of Sections D and 5.4 does not depend on between-model variance.

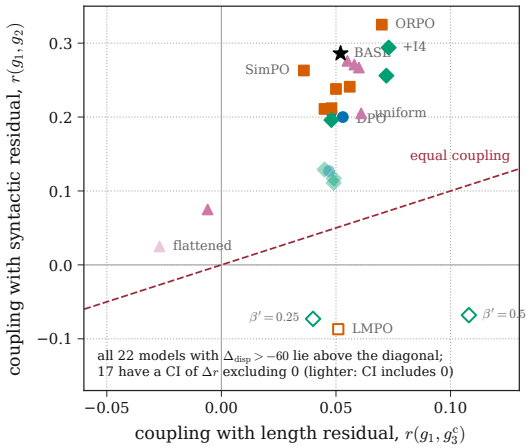


Figure 3: Within-model coupling of the semantic gap with the syntactic residual (y) against its coupling with the length residual (x), Pearson over the 1,979 families, one marker per model (colours as in Fig. 1). Every model except the three with large displacement lies above the diagonal; lighter markers have a Δr interval that contains zero.

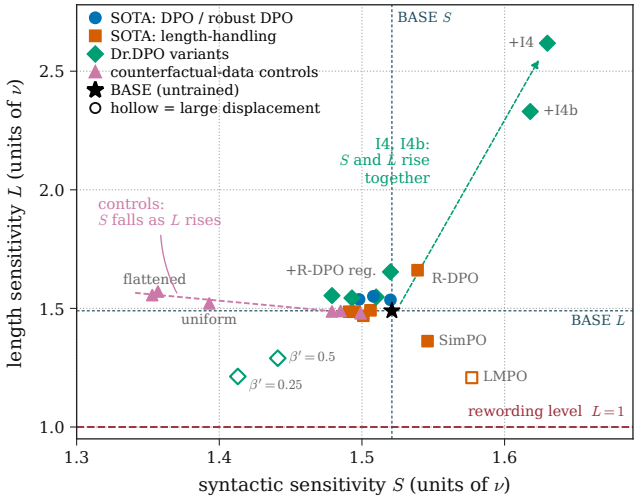


Figure 4: Syntactic against length sensitivity for the 25 models (Section 5.2). The surface components co-vary in every training regime, differently in each: they trade off in the controlled group, rise together under mean-length penalties, and the two sharp-DRO Dr. DPO variants ($\beta' \leq 0.5$), whose twin scale is inflated, fall on both.

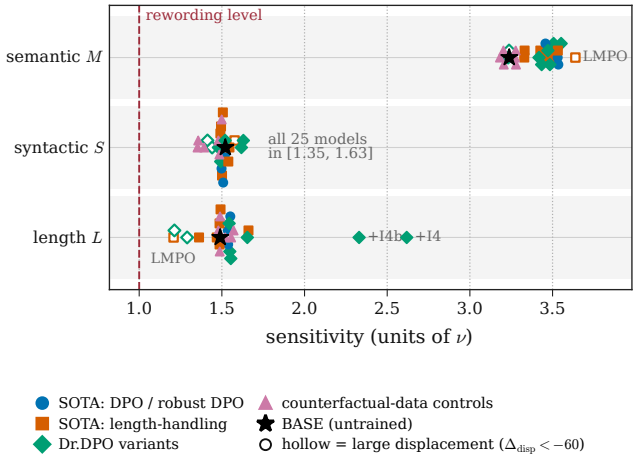


Figure 5: M, S and L of the 25 models as strips, in units of ν. S stays in a narrow band around 1.5, where the untrained model sits, whereas L spreads in both directions.

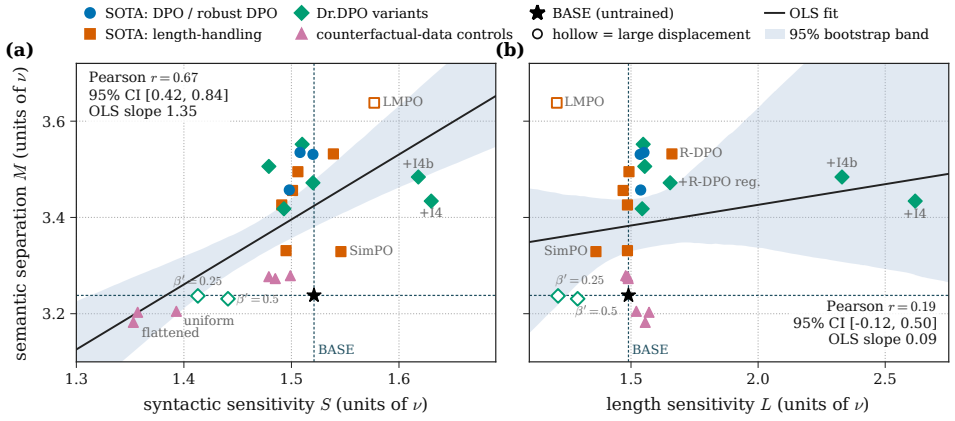


Figure 6: Cross-model view, for context only: M against S (left) and L (right), one marker per model, coloured by training regime (large displacement hollow, BASE a star). The correlation with S reflects regime structure; the per-family analysis of Section 5.4 is the relevant test. The per-model coupling contrast is in Fig. 3.

F ADDITIONAL EXPERIMENTS

All analyses follow the protocol of Section B.4 and use the configurations of Table 7; checkpoints, score files and scripts are released with the data. In brief: further training seeds, a penalty-strength sweep, a pair-type ablation, independently drawn twins and a second backbone leave the findings of Section 5 in place.

F.1 TRAINING SEEDS

Two further seeds of DPO and SimPO (Table 13) keep every run's S within 0.03 of the untrained 1.521. DPO $>$ BASE on C and $\bar{g}_1$ holds for all three seeds and on the seed mean ($+0.021$ $[+0.007, +0.035]$ and $+0.378$ $[+0.289, +0.462]$); SimPO $<$ DPO holds on $\bar{g}_1$ in all nine seed pairings and on the means (-0.246 $[-0.317, -0.173]$), and on C in six of nine pairings and on the means (-0.018 $[-0.028, -0.007]$).

Table 13: Training-seed replication of DPO and SimPO (identical configurations; seeds 1 and 2 added to the seed-42 checkpoints of Table 2). Seed standard deviations are at most $0.27\times$ the family-bootstrap CI half-width for every DPO statistic and $0.53\times$ for SimPO's C.

Run	acc	M	S	L	C	$\bar{g}_1$	Δ_{disp}
DPO, seed 42 (Table 2)	0.580	3.535	1.508	1.551	0.790	2.73	-14.7
DPO, seed 1	0.581	3.541	1.515	1.509	0.799	2.77	-11.3
DPO, seed 2	0.584	3.512	1.512	1.532	0.794	2.72	-13.6
DPO, mean	0.581	3.529	1.512	1.531	0.794	2.74	-13.2
DPO, seed sd	0.002	0.015	0.004	0.021	0.005	0.03	
SimPO, seed 42 (Table 2)	0.511	3.329	1.546	1.362	0.771	2.44	-0.5
SimPO, seed 1	0.524	3.362	1.530	1.413	0.788	2.58	$+6.6$
SimPO, seed 2	0.516	3.318	1.538	1.356	0.772	2.46	-0.2
SimPO, mean	0.517	3.336	1.538	1.377	0.777	2.49	$+2.0$
SimPO, seed sd	0.007	0.023	0.008	0.032	0.010	0.07	

F.2 PENALTY-STRENGTH SWEEP

R-DPO ($\alpha_0 = 0.05$) and LD-DPO ($\alpha_0 = 0.5$) were retrained with four further values of their length parameter each, everything else unchanged (Table 14). Paired against DPO, C is indistinguishable at all ten levels ($|\Delta C| \leq 0.007$) and $\bar{g}_1$ at nine (LD-DPO $\alpha = 0.2$: -0.067 $[-0.117, -0.014]$). R-DPO's L exceeds DPO's from $\alpha = 0.025$ on ($+0.766$ $[+0.696, +0.839]$ at $\alpha = 0.2$) and its S from $\alpha = 0.05$ on ($+0.122$ $[+0.078, +0.167]$ at $\alpha = 0.2$): at four times the strength of Table 2 the length penalty leaks into the syntactic channel. Across the five R-DPO levels Spearman$(\alpha, L) = +1$ and Spearman$(L, \text{acc}) = +1$, while Spearman$(L, C) = -0.8$; LD-DPO's L moves by 0.06 over $\alpha \in [0.2, 0.8]$ and its accuracy not at all.

F.3 PAIR-TYPE ABLATION

Plain DPO was trained on the 1,998 prompts with one pair type per arm and the optimizer settings of the control arms (Table 15). The two matched arms are indistinguishable on C, S, L and acc and differ only slightly on $\bar{g}_1$ ($+0.024$ $[+0.004, +0.043]$ for the syntax-matched arm); both exceed the original-pair arm on C and acc, and both raise L by 0.06. As for the flattened controls, the fall of S in units of ν comes with a rise of the twin scale: $\mathbb{E}|g_2|$ in raw log-probability rises from 12.9 to 15.5 and 17.3 while $\bar{\nu}$ rises from 8.6 to 11.4 and 12.8. The step-matched flattened arm, which sees all three views for the same number of steps, reaches $C = 0.820$; the $3\times$-step arm 0.834. Matching length as text is what helps, and it helps as much as matching the sentence pattern on top of it: the operative surface difference is the one that travels with length, which a penalty on the token count leaves untouched (Section F.2). What these arms do not separate is matching from rewriting: every matched arm also trains against LLM-written negatives, so a control trained on unconstrained meaning-preserving paraphrases of the rejected summary, with the same prompts, seeds and steps, is the comparison that would tell the two apart; it is not part of this version.

Table 14: Penalty-strength sweep on the 92,858-pair split (same recipe and seed as Table 2; only α varies). R-DPO's α moves L monotonically over a 16× range and, at $\alpha \geq 0.05$, raises S; LD-DPO is nearly inert. C is indistinguishable from DPO at all ten levels (paired $|\Delta C| \leq 0.007$, every CI contains zero).

Model	acc	S	L [CI]	C [CI]	$\bar{g}_1$	Δ_{disp}
R-DPO α=0.0125	0.549	1.528	1.551 [1.467, 1.633]	0.795 [0.777, 0.812]	2.70	−6.7
R-DPO α=0.025	0.557	1.530	1.585 [1.499, 1.669]	0.797 [0.779, 0.814]	2.71	−6.8
R-DPO α=0.05 (Table 2)	0.565	1.539	1.661 [1.570, 1.751]	0.793 [0.774, 0.810]	2.74	−5.8
R-DPO α=0.1	0.574	1.555	1.847 [1.747, 1.951]	0.793 [0.775, 0.811]	2.76	−7.6
R-DPO α=0.2	0.586	1.630	2.317 [2.197, 2.442]	0.790 [0.772, 0.807]	2.71	−13.0
LD-DPO α=0.2	0.549	1.493	1.458 [1.383, 1.536]	0.794 [0.776, 0.811]	2.66	−12.6
LD-DPO α=0.4	0.550	1.502	1.481 [1.404, 1.560]	0.793 [0.776, 0.811]	2.68	−11.7
LD-DPO α=0.5 (Table 2)	0.548	1.506	1.492 [1.414, 1.571]	0.794 [0.776, 0.811]	2.69	−11.0
LD-DPO α=0.6	0.550	1.512	1.500 [1.422, 1.579]	0.793 [0.776, 0.811]	2.69	−10.1
LD-DPO α=0.8	0.549	1.517	1.517 [1.438, 1.597]	0.794 [0.776, 0.811]	2.69	−8.9
BASE	0.518	1.521	1.490 [1.418, 1.565]	0.774 [0.754, 0.792]	2.36	+0.0
DPO ($\beta = 0.05$)	0.580	1.508	1.551 [1.469, 1.637]	0.790 [0.771, 0.807]	2.73	−14.7

Table 15: Pair-type ablation on the 1,998-prompt set: plain DPO ($\beta = 0.1$, 250 steps, three seeds; seed means) trained on one pair type per arm. Paired same-family contrasts (95% bootstrap CIs): length-matched − original: C +0.025 [+0.015, +0.035], acc +0.013 [+0.004, +0.024]; length- and syntax-matched − original: C +0.030 [+0.019, +0.041], acc +0.016 [+0.006, +0.027]; syntax-matched − length-matched: C +0.005 [−0.002, +0.011], S −0.006 [−0.019, +0.007], $\bar{g}_1$ +0.024 [+0.004, +0.043]. The twin scale $\bar{\nu}$ (raw log-probability) rises when the model is trained against rewrites, so S falls in units of ν while $\mathbb{E}|g_2|$ rises from 12.9 to 15.5 and 17.3.

Training pair	acc	M	S	L	C	$\bar{g}_1$	$\bar{\nu}$	Δ_{disp}
original pair (y_w, y_l)	0.518	3.273	1.501	1.482	0.775	2.41	8.6	+0.1
length-matched (y_w, y_1'')	0.532	3.120	1.355	1.548	0.801	2.45	11.4	−6.8
length- and syntax-matched (y_w, y_1')	0.535	3.101	1.348	1.541	0.805	2.47	12.8	−9.5

F.4 INDEPENDENTLY DRAWN TWINS

To test whether the ordering bias of the twins (Section 3.3) is a co-generation effect, we regenerated both twins of every family as separate single-output writer calls, each audited alone, under the original specification with the same model and checks plus a vocabulary-overlap threshold (3,958 calls; 1,899 of the 1,979 families pass both), keeping the path rewrites y_1', y_1'' fixed, and re-scored only the new twins with all 37 checkpoints. For the length twin the bias disappears: $\mathbb{E}[B_1 - B_{\text{ind}}]$ lies between -0.21 and $+0.75$ raw log-probability with an interval containing zero for each of the 22 models with ordinary displacement, and $\Pr(B_1 > B_{\text{ind}}) = 0.49$–$0.51$ (against +4.5–10.8 and 0.66–0.71 for the co-written twin). For the syntax twin it persists ($\mathbb{E}[A_1 - A_{\text{ind}}] = +3.2$–$6.4$, $\Pr = 0.58$–0.61), because y_1' and y_1'' were written together whereas the new twin is derived from the fixed y_1''; separating this construction effect from an ordering effect would require regenerating both versions of both rewrites. With the independent twins $\bar{\nu}$ rises by 4–13%, S stays above 1 for every model (smallest lower confidence bound 1.224), $\hat{d}$ remains positive for all 22, and the rankings by L and M are preserved ($\rho_s = 0.86$) while the S ranking is not (0.57).

F.5 SECOND BACKBONE: LLAMA-3.2-3B-INSTRUCT

We repeated the protocol on Llama-3.2-3B-Instruct with every Qwen configuration of Table 7 unchanged except the model: the untrained instruct model again serves as reference and initialization; nine SOTA methods (Robust-DPO and the Dr. DPO variants were not replicated) and the six controls with three seeds, 28 checkpoints in all. Because the token-exact construction is specific to the Qwen tokenizer, every Llama gap is the content component of Eq. (8); the rewrites remain token-exact

under the Llama tokenizer in 61.4% of families, on which raw and content gaps coincide and every result below holds. The untrained model has $S = 1.520$ and $\bar{\nu} = 8.9$ against 1.521 and 8.7 on Qwen, and the first-version twin bias appears with the same size ($+4.2$ and $+4.5$), so it is a property of the data. Seven predictions were written down before any trained Llama checkpoint was scored; six hold: across the SOTA methods with ordinary displacement S varies nearly four times less than L, which moves in both directions; no length-handling objective raises C or $\bar{g}_1$ above DPO, and five are significantly below it on $\bar{g}_1$; the flattened arms raise C over the original-pair control by $+0.040$ $[+0.029, +0.053]$ and $+0.035$ $[+0.023, +0.047]$; $S \geq 1$ for all 16 models; accuracy follows L ($\rho_s = 0.62$) slightly more than S (0.58); and the rankings of the ten shared models by acc, L and C transfer across backbones ($\rho_s = 0.87$, 0.85, 0.70) while the S ranking does not (0.13), as expected for a statistic with little between-model variance. The prediction that every length-handling objective keeps S within 0.05 of the untrained value fails for ORPO (-0.078). Three Llama models show large displacement: SimPO, LMPO and the $3\times$-step flattened arm. The per-family association of Section 5.4 is weaker on Llama, with the contrast $\mathrm{corr}(g_1, g_2) - \mathrm{corr}(g_1, g_3^c)$ significantly positive in 7 of the 13 models with ordinary displacement (2 of 13 on the token-exact subset), and the significance class of the 24 contrast–outcome entries of Table 17 agrees with Qwen in 21.

Table 16: Second backbone: Llama-3.2-3B-Instruct trained with the Qwen configurations of Table 7 unchanged (16 models, 28 checkpoints), scored on the same 1,979 families with content-component gaps (Eq. (8)). Last column: the same configuration's S / L / C on Qwen2.5-1.5B-Instruct. †Large displacement ($\Delta_{\mathrm{disp}} < -60$).

Model	acc	M	S	L	C	$\bar{g}_1$	$\bar{\nu}$	Δ_{disp}	Qwen S / L / C
BASE	0.520	3.066	1.520	1.423	0.780	2.30	8.9	$+0.0$	1.521 / 1.490 / 0.774
SOTA methods (92,858 pairs)									
DPO	0.626	3.786	1.504	1.640	0.804	3.07	12.5	-10.3	1.508 / 1.551 / 0.790
R-DPO	0.604	3.619	1.533	1.734	0.810	2.93	10.1	-3.6	1.539 / 1.661 / 0.793
LD-DPO	0.598	3.576	1.496	1.469	0.809	2.88	10.8	-10.1	1.506 / 1.492 / 0.794
SamPO	0.588	3.567	1.506	1.424	0.806	2.86	10.9	-16.4	1.501 / 1.469 / 0.792
SimPO†	0.585	4.059	1.486	1.298	0.762	3.00	9.2	-106.8	1.546 / 1.362 / 0.771
ORPO	0.519	3.332	1.442	1.495	0.806	2.67	8.4	$+35.0$	1.495 / 1.486 / 0.784
Tie training	0.581	3.548	1.480	1.477	0.811	2.86	9.8	$+0.9$	1.491 / 1.486 / 0.789
LMPO†	0.582	3.991	1.476	1.242	0.763	2.96	9.2	-98.5	1.577 / 1.208 / 0.754
Dr. DPO	0.656	3.597	1.528	1.638	0.796	2.88	19.9	-42.7	1.498 / 1.538 / 0.790
Counterfactual-data controls (1,998 prompts, 3 seeds)									
Dr. DPO, original pairs	0.547	3.235	1.488	1.436	0.791	2.52	9.7	-8.9	1.499 / 1.479 / 0.775
DPO, KL-DRO views	0.534	3.178	1.477	1.419	0.795	2.48	9.7	-2.5	1.485 / 1.489 / 0.781
Dr. DPO, KL-DRO views	0.547	3.217	1.445	1.420	0.800	2.55	10.4	-9.1	1.479 / 1.487 / 0.781
Dr. DPO, uniform views	0.546	3.166	1.374	1.456	0.813	2.61	12.8	-11.6	1.393 / 1.521 / 0.796
flattened, $3\times$ steps†	0.574	3.169	1.333	1.465	0.832	2.69	22.4	-68.7	1.357 / 1.571 / 0.834
flattened, step-matched	0.558	3.129	1.323	1.466	0.826	2.63	19.8	-55.3	1.353 / 1.556 / 0.820

Table 17: The paired contrasts of Table 8 on Llama-3.2-3B-Instruct next to Qwen2.5-1.5B-Instruct (same families; 95% bootstrap CIs; controls seed-averaged). The significance class agrees in 21 of the 24 entries; on Llama five length-handling objectives fall significantly below DPO on $\bar{g}_1$. †Large displacement on Llama.

Contrast		Llama-3.2-3B	Qwen2.5-1.5B
DPO − BASE	C	$+0.024$ $[+0.009, +0.040]$	$+0.016$ $[+0.002, +0.031]$
	$\bar{g}_1$	$+0.764$ $[+0.655, +0.882]$	$+0.367$ $[+0.277, +0.452]$
R-DPO − DPO	C	$+0.006$ $[-0.005, +0.017]$	$+0.003$ $[-0.007, +0.014]$
	$\bar{g}_1$	-0.140 $[-0.205, -0.080]$	$+0.012$ $[-0.035, +0.061]$
LD-DPO − DPO	C	$+0.005$ $[-0.005, +0.016]$	$+0.004$ $[-0.006, +0.014]$
	$\bar{g}_1$	-0.188 $[-0.251, -0.130]$	-0.041 $[-0.092, +0.013]$
SamPO − DPO	C	$+0.002$ $[-0.010, +0.012]$	$+0.002$ $[-0.008, +0.013]$
	$\bar{g}_1$	-0.203 $[-0.271, -0.139]$	-0.088 $[-0.144, -0.030]$
SimPO† − DPO	C	-0.042 $[-0.059, -0.025]$	-0.019 $[-0.031, -0.007]$
	$\bar{g}_1$	-0.071 $[-0.241, +0.099]$	-0.290 $[-0.367, -0.214]$
ORPO − DPO	C	$+0.002$ $[-0.015, +0.018]$	-0.006 $[-0.020, +0.009]$
	$\bar{g}_1$	-0.396 $[-0.530, -0.272]$	-0.206 $[-0.294, -0.120]$
Tie training − DPO	C	$+0.007$ $[-0.005, +0.018]$	-0.001 $[-0.012, +0.011]$
	$\bar{g}_1$	-0.208 $[-0.289, -0.135]$	-0.097 $[-0.163, -0.029]$
LMPO† − DPO	C	-0.042 $[-0.059, -0.024]$	-0.036 $[-0.056, -0.016]$
	$\bar{g}_1$	-0.111 $[-0.273, +0.055]$	-0.163 $[-0.342, +0.017]$
flattened $(3\times)^{\dagger}$ − original pairs	C	$+0.040$ $[+0.029, +0.053]$	$+0.059$ $[+0.044, +0.073]$
	$\bar{g}_1$	$+0.179$ $[+0.095, +0.266]$	$+0.301$ $[+0.205, +0.396]$
flattened (step-m.) − original pairs	C	$+0.035$ $[+0.023, +0.047]$	$+0.045$ $[+0.032, +0.058]$
	$\bar{g}_1$	$+0.119$ $[+0.042, +0.193]$	$+0.215$ $[+0.140, +0.294]$
flattened $(3\times)^{\dagger}$ − Dr. DPO, KL-DRO	C	$+0.032$ $[+0.021, +0.043]$	$+0.053$ $[+0.040, +0.068]$
	$\bar{g}_1$	$+0.144$ $[+0.065, +0.224]$	$+0.272$ $[+0.178, +0.364]$
flattened (step-m.) − Dr. DPO, KL-DRO	C	$+0.026$ $[+0.015, +0.038]$	$+0.039$ $[+0.027, +0.052]$
	$\bar{g}_1$	$+0.084$ $[+0.017, +0.150]$	$+0.186$ $[+0.114, +0.261]$