GRADED EQUIVALENCE CERTIFICATES FOR PREFERENCE OPTIMISATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Preference optimisation has dozens of objectives published as distinct methods, and several frameworks that place them in one parameterised family. Family membership does not tell a practitioner whether switching methods changes anything. Every objective we study gives each response a *score*, takes the gap between two scores as a *margin*, and applies a *link*, a one-dimensional loss, to the margin. We ask at which level two objectives are the same and separate four. Two objectives can share the ranking they induce over responses, the score up to a per-prompt scale and shift, the loss, or the policy that training converges to. Each level has a certificate, an algebraic check that two objectives agree, and a witness, a few numbers that show they do not. The first three levels form a strict chain. The fourth follows from a shared loss and from nothing weaker. The optimum is set by the loss's *target map*, the score gap the loss drives a labelled pair towards as a function of how often annotators prefer one response. Methods that share DPO's score have different target maps and converge to different policies. Applying the certificates to the methods in current use shows that a shared loss is rare and a shared ranking is common. On a trained model the certified relationships hold exactly, and four links that share DPO's score order their margins as their target maps predict. We also measure the gap the certificates leave open, between scores that use a reference model and scores that use the policy alone. A closed-form experiment shows that when the labelled pairs form a chain, the choice of link alone can reverse which response the policy prefers, and that comparing every response against one anchor removes the effect for every link.

1 INTRODUCTION

Reinforcement learning from human feedback (Christiano et al., 2017; Ouyang et al., 2022) trains a language model in two stages: fit a reward model to human preference labels, then optimise the policy against it with PPO. DPO (Rafailov et al., 2023) collapsed the two stages into a single loss on preference pairs. That loss now has dozens of successors: IPO (Azar et al., 2024), SimPO (Meng et al., 2024), ORPO (Hong et al., 2024), KTO (Ethayarajh et al., 2024), a game-theoretic family (SPPO, Nash-MD, INPO, DNO) and a group-normalised policy-gradient family led by GRPO (Shao et al., 2024). Each is published as a distinct method. A practitioner who switches from one to another wants to know what the switch changes. It may change the policy the model is pushed towards, the loss surface it is pushed along, or only the notation.

ΨPO (Azar et al., 2024), GPO (Tang et al., 2024b), RPO (Sun et al., 2025), RainbowPO (Zhao et al., 2025) and the taxonomy of Raheja & Pochhi (2026) each place several methods inside one template. Membership in a template shows that two methods can be written in one notation. Whether they are the same is a further question, and it has no answer until one says at which level sameness is demanded. This paper poses that question and answers it in four parts.

Question 1. *When are two preference-optimisation objectives the same?*

At one of four levels. Every method we study gives each response a number, its *score*. The *margin* is the gap between the two scores in a labelled pair, and a *link* turns the margin into a loss. Two methods can agree on which response scores higher (Rank), on the score itself up to a per-prompt scale and

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

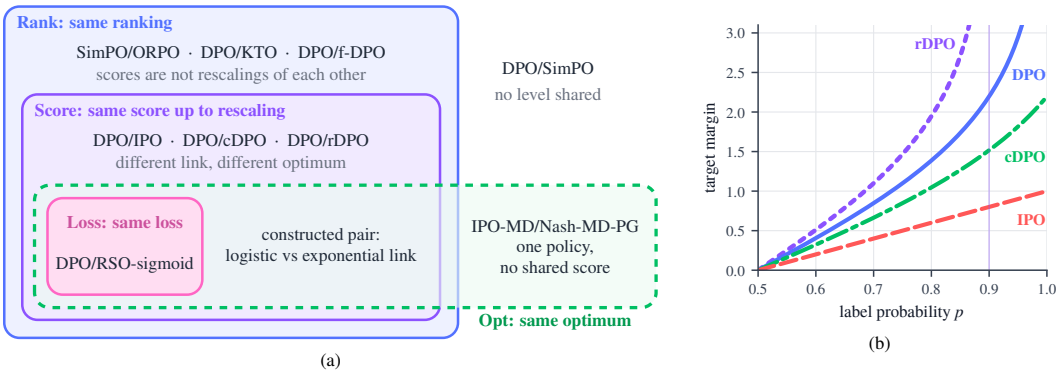

Figure 1: (a) The four levels of Section 3. Loss ⊂ Score ⊂ Rank are nested, and Opt cuts across them. The logistic/exponential pair is a constructed example (Proposition 1(d)). (b) Target maps of the four losses that share DPO's score. Four maps give four optima.

shift (Score), on the whole loss (Loss), or on the policy training converges to (Opt). Section 3 defines the levels, shows that none of the implications between them reverses (Figure 1a), and gives each level a *certificate* that confirms it and a *witness* that refutes it.

Question 2. *Two methods share a score. Do they share an optimum?*

No. The optimum is decided by the link through its *target map*, the margin the link aims for on a pair that annotators prefer with a given probability (Section 4). DPO, IPO, cDPO (DPO with label smoothing) and rDPO (DPO corrected for label noise) share one score and have four target maps. On a pair that annotators prefer 80% of the time, they aim for margins that differ by a factor of three at common settings (Figure 1b). On a trained model the four order their margins as their target maps predict (Experiment 3, Section 6).

Question 3. *How are the methods in use actually related?*

Mostly by ranking. We took the training objective of 33 methods from each paper's own source and ran the certificates (Section 5). Two losses are shared by more than one method, and in both cases the sharing methods differ only in their training data. Shared rankings are common and cross published families.

Question 4. *How large are the gaps the certificates cannot measure, and does the choice of labelled pairs matter?*

On a trained 0.5B model, scores that use the reference model and scores that use the policy alone agree on which of two responses is better for 57% to 63% of response pairs (Experiment 1, Section 6). In a closed-form experiment, labelling pairs along a chain lets the link reverse a preference that annotators hold 88% of the time, and labelling every response against one anchor prevents it (Experiment 2, Section 6).

2 What a preference-optimisation objective is made of

Most methods in this paper build their loss in three steps. Naming the steps gives all the notation the main text needs (Table 1).

Score, margin, link. First, the method gives each response y to a prompt x a single number, its *score*. DPO's score is $\beta \log(\pi_\theta(y \mid x)/\pi_{\text{ref}}(y \mid x))$. It measures how much more likely the trained policy π_θ makes y than the frozen reference model π_{ref} does. The coefficient β sets how strongly training holds the policy to the reference. Second, it subtracts the scores of the two responses in a labelled pair to get the *margin* $\Delta = s_\theta(x, y) - s_\theta(x, y')$. Third, a *link* ℓ turns the margin into a loss. DPO's link $-\log \sigma(\Delta)$, with σ the sigmoid, is large when the preferred response scores lower and falls towards zero as the margin grows. The link is where a method says how hard to push a pair apart,

Table 1: Three methods as score, link and weight. τ and γ are IPO's and SimPO's own parameters.

Method	Score $s_\theta(x, y)$	Link $\ell(\Delta)$	Weight ω
DPO (Rafailov et al., 2023)	$\beta \log \big(\pi_\theta(y \mid x)/\pi_{\mathrm{ref}}(y \mid x)\big)$	$-\log \sigma(\Delta)$	1
IPO (Azar et al., 2024)	$\log \big(\pi_\theta(y \mid x)/\pi_{\mathrm{ref}}(y \mid x)\big)$	$(\Delta - \frac{1}{2\tau})^2$	1
SimPO (Meng et al., 2024)	$\frac{\beta}{\lvert y\rvert} \log \pi_\theta(y \mid x)$	$-\log \sigma(\Delta - \gamma)$	1

and Section 4 makes this precise. Reid & Williamson (2010) use the same word for the map from a probability to a prediction. That map is our target map, and Section 8 gives the correspondence. A pair may also carry a weight ω, and the training loss is the weighted average over the data $\mathcal{D}$,

$$\mathcal{L}(\theta) = \mathbb{E}_{(x,y,y')\sim\mathcal{D}}\big[\,\omega\,\ell\big(\Delta_\theta(x, y, y')\big)\big]. \tag{1}$$

With a soft label, where the first response wins with probability p, a pair contributes $p\,\ell(\Delta) + (1 - p)\,\ell(-\Delta)$. We call a loss of this form a *margin objective*. Scores like DPO's, which compare the policy with the reference model, are *reference-based*. Scores that use the policy's own likelihood alone, like SimPO's and ORPO's, are *reference-free*.

Objectives outside this form. *Per-response objectives* (KTO, BCO, Anchored PO, SPPO) score each response on its own against a fixed reference point. *Policy-gradient objectives* (GRPO, RLOO, PPO, ReMax, Nash-MD-PG) weight each sampled response by an advantage. For every method we write a *loss card* with five fields: the score, the link, the weight, any extra terms outside the margin (such as an SFT loss), and the data the loss is trained on.

3 Four levels of "the same"

What can it mean for two objectives to be the same? One of four things, from weakest to strongest (Table 2). Think of two judges marking the same essays. They agree at Rank if they always pick the same winner. They agree at Score if one judge's marks are the other's after a change of units. They agree at Loss if they also penalise a wrong call in the same way. They agree at Opt if, trained on the same cases, they end up making the same decisions. The formal definitions are in Appendix A.

Table 2: The four levels, each with a pair that agrees and a pair that does not.

Level	Two objectives agree on	Agree	Do not agree
Rank	which of two responses scores higher, for every parameter setting	SimPO, ORPO	DPO, SimPO
Score	the score itself, up to a per-prompt scale and shift	DPO, IPO	SimPO, ORPO
Loss	the whole loss, up to a constant factor	DPO, RSO-sigmoid	DPO, IPO
Opt	the policy they converge to on one dataset	IPO-MD, Nash-MD-PG	DPO, IPO

How the levels relate. The same loss gives the same score, and the same score gives the same ranking. None of these implications reverses. The Score and Loss rows of Table 2 give a pair that breaks each converse. Sharing an optimum sits outside this chain. Two objectives with one loss share an optimum on any dataset. Objectives with different losses can share one too (Figure 1a), and an optimum shared on one dataset says nothing about prompts outside it. Proofs are in Appendix B.

How does one decide the level for a given pair of methods? Mechanically, once each method's loss card is written down. The card is transcribed by hand from the method's source equation. This transcription is the one step a reader must trust, and Appendix D documents it equation by equation. The scale and shift in Score are allowed to depend on the prompt because responses are only ever compared within one prompt, where such a change leaves the winner unchanged. The score and loss levels are decided on the loss-card formulas by normalising each score per prompt (subtract the mean over the prompt's responses, divide by the spread), comparing what remains, and checking that the ratio of the two spreads does not depend on the parameters. For DPO and IPO this gives one normal form, since IPO's score is DPO's divided by the constant β. The rank level is confirmed by an increasing function that maps one score to the other. ORPO's score, for example, is an increasing function of SimPO's that is not of the form $a\,s + b$. The rank level is refuted by two pairs that the

methods order differently. The optimum level needs a theorem (Section 4). Whenever a level fails, the output is a *witness*, a few numbers anyone can check by hand. For SimPO and ORPO the witness is two response pairs under one prompt with the same SimPO margin, 0.1, and ORPO margins of 0.744 and 0.105 (the two pairs are in Appendix D). No per-prompt scale and shift maps one score to the other. Appendix B.1 lists one witness type for each way a level can fail.

What does this look like on one prompt? Table 3 runs five scores on one prompt with two responses. The trained policy gives the preferred response y log-likelihood -10 over 5 tokens and the rejected y' log-likelihood -12 over 4 tokens. The reference model gives both -11. Reading down the table decides three levels by hand. DPO's and IPO's scores are the same numbers up to the factor $\beta = 0.1$, so their margins differ by that factor on every prompt and the pair passes Score. ORPO's score is $\mathrm{logit} \circ \exp$ of SimPO's score divided by β. That map is increasing, so the two always agree on the winner and pass Rank. The map is not affine, which the witness above shows, so Score fails. DPO and SimPO fail Rank as soon as the reference model has its own opinion. Change the reference log-likelihoods to -8 for y and -13 for y', and give y 10 tokens. DPO now scores y at -0.2 and y' at $+0.1$, so DPO prefers y'. SimPO scores them -2 and -6, so SimPO prefers y. Those two lines are the whole witness. The table leaves one level open, Loss for DPO against IPO. The links differ, so the question is what each link does with the margin it is handed. The next section answers it.

Table 3: Five scores on one prompt ($\beta_{\mathrm{DPO}} = 0.1$, $\beta_{\mathrm{SimPO}} = 2$, $\mathrm{logit}\, P = \log \frac{P}{1-P}$). The last column names the level the row pair certifies.

Score	Formula	$s(y)$	$s(y')$	Δ	Reads as
DPO	$\beta \log(\pi_\theta/\pi_{\mathrm{ref}})$	0.10	-0.10	0.20	Score: IPO $=$ DPO$/\beta$
IPO	$\log(\pi_\theta/\pi_{\mathrm{ref}})$	1.00	-1.00	2.00	
SimPO	$(\beta/\lvert y\rvert) \log \pi_\theta$	-4.00	-6.00	2.00	Rank only: $\mathrm{logit} \circ \exp$ is increasing, not affine
ORPO	$\mathrm{logit} \exp(\log \pi_\theta/\lvert y\rvert)$	-1.86	-2.95	1.09	
SLiC	$\log \pi_\theta$	-10.00	-12.00	2.00	no reference, no length normalisation

4 What a shared score does and does not transfer

Two methods have the same score. What else do they share? Their rankings at every point in training. Whether they end at the same trained policy depends on the link, and one curve per link decides it.

The target map. Take a single pair whose label says the first response wins with probability p, and minimise one link's loss on that pair alone. The margin where the loss is smallest is the link's *target*, and the curve of targets against p is its *target map*. It answers a concrete question. Given how often annotators prefer one response, how far apart does this loss want the two scores to be? Table 4 lists the map of every link in use, and Figure 1b plots the four that share DPO's score. At $p = 0.8$ those four want margins from 0.60 (IPO) to 1.95 (rDPO).

Result 1. *Two margin objectives with the same score reach the same optimum exactly when their target maps are proportional*, with the factor equal to the scale between the two scores. The factor is needed because Score lets one score be a rescaled copy of the other. The formal version (Theorem 1) assumes the model is flexible enough to place every labelled pair at its target, which also needs the labelled pairs of a prompt not to close a cycle. A chain or a star qualifies. When every pair is labelled the targets can conflict and the minimiser is a compromise, which Proposition 4(c) treats. DPO, IPO, cDPO and rDPO have four non-proportional target maps, so on such a design, with labels spread over an interval, they train towards four different policies (Corollary 1 states the exact condition). With hard 0/1 labels the difference is starker. DPO and rDPO push margins without bound. IPO and cDPO stop at a finite margin set by their parameters (Table 4).

Monotone links are the wrong condition. A natural guess is that any two links that fall as the margin grows give the same optimum. The logistic and exponential links both fall, and their target maps are $\mathrm{logit}\, p$ and $\frac{1}{2}\,\mathrm{logit}\, p$, so their optima differ. The squared link rises again beyond its target and its target map rises with p (Proposition 5).

Which changes to a score are free? Adding a per-prompt constant to every score changes nothing, because it cancels in every margin. Multiplying the score for one prompt by a positive factor keeps

Table 4: Target map of each link, the margin it drives a pair towards. $\operatorname{logit} p = \log \frac{p}{1-p}$; values at $p = 0.8$ use $m = 1$, $\epsilon = 0.1$, $\gamma = \frac{1}{2}$, and the gaps between links depend on these settings.

Link	Target map	p=0.8	Used by
logistic	$\operatorname{logit} p$	1.39	DPO, ORPO, CPO, R-DPO, WPO, RSO, f-DPO
squared, target m	$m(2p-1)$	0.60	IPO ($m = \frac{1}{2\tau}$), IPO-MD, INPO
label-smoothed, ϵ	$\operatorname{logit}(\epsilon + (1-2\epsilon)p)$	1.05	cDPO, RankNet (soft labels)
noise-corrected, ϵ	$\operatorname{logit} \frac{p-\epsilon}{1-2\epsilon}$	1.95	rDPO
shifted logistic, γ	no closed form	1.73	SimPO, α-DPO
exponential	$\frac{1}{2} \operatorname{logit} p$	0.69	(reference)
hinge, margin m	$m \operatorname{sign}(2p-1)$, a step	1.00	SLiC-HF, RSO-hinge, RRHF
soft-label logistic	set per pair	—	DNO

every ranking and moves the optimum. DPO's score is β times an implicit reward r, and the KL-regularised optimum is $\pi^* \propto \pi_{\text{ref}} \exp(r/\beta)$. Scaling r on one prompt is the same as changing β on that prompt, so π^* moves (Proposition 3). This is why a shared score does not imply a shared optimum. GRPO samples a group of responses per prompt and divides each reward, after centring, by the group's standard deviation. That divisor is exactly such a factor (Corollary 2). Call the difference of two responses' advantages their *effective margin*. A pair with rewards 1.0 and 0.6 has effective margin 2.00 in a group of two and 0.91 in a group of four with rewards $(1.0, 0.6, 0.3, -0.2)$ (Appendix B.2). Batch composition therefore rescales GRPO's updates and leaves their order unchanged. In a group of two the advantages are ± 1 whatever the reward gap, so the update carries only which response won. This is the regime in which Wu et al. (2025) find that two samples retain most of the performance of sixteen.

5 Thirty-three methods, from source

How are the methods in use related? Table 5 gives the answer for 33 methods, and Appendix D derives each row from the paper's own source equation. A script that prints what it computed checked every equivalence and every witness. Where two methods share an acronym, the objective decides which one a row means.

What the classification shows.

- *Same loss is rare.* The 22 losses of the 21 margin objectives (RSO defines two) fall into 20 loss classes. The two shared classes are DPO with RSO-sigmoid and IPO with IPO-MD. Each is one loss trained on different data.
- *Same ranking is common and crosses published families.* DPO's ranking is shared by nine other margin losses and by the per-response scores of KTO, BCO and Anchored PO. The nine include f-DPO, which replaces DPO's KL term with an f-divergence, for every convex f. SimPO, RRHF and ORPO share a ranking and differ at every level above it.
- *Twelve of the 33 have no pairwise margin.* Four score responses one at a time, six are policy-gradient or token-level methods, and two are listwise. SPPO is among the four. Two training states that give a pair the same margin have SPPO losses of 0.5 and 12.5, so its loss is not a function of the margin (Proposition 6).
- *Extra terms and unequal weights break the loss level.* ORPO, CPO, RRHF, SLiC-HF and XPO add a term outside the margin. WPO, LambdaRank, α-DPO and DNO weight pairs unequally. In both cases only the margin part can be compared.

6 Experiments

Experiment 1: the levels on a trained model

Do the certified relationships hold in a trained model, and how large are the gaps the certificates leave open?

Table 5: The 33 methods. A shared letter in the Rank or Loss column means equivalence at that level. Subscript t marks a class that depends on the training iterate, and f one class per divergence. RSO has two rows. Formulas are in Table 10 and derivations in Appendix D.

Method	Score	Link	Pair weight	Reference	Extra term	Rank	Loss
Margin objectives							
DPO	log-ratio to reference	logistic	equal	fixed		A	A1
RSO (sigmoid)	log-ratio to SFT model	logistic	equal	fixed		A	A1
RSO (hinge)	log-ratio to SFT model	hinge, m=1	equal	fixed		A	A2
cDPO	log-ratio to reference	label-smoothed	equal	fixed		A	A3
rDPO	log-ratio to reference	noise-corrected	equal	fixed		A	A4
IPO	log-ratio to reference	squared	equal	fixed		A	A5
IPO-MD	log-ratio to reference	squared	equal	fixed		A	A5
WPO	log-ratio to reference	logistic	pair	fixed		A	A6
XPO	log-ratio to reference	logistic	equal	fixed	exploration	A	A7
f-DPO	f' of the ratio to reference	logistic	equal	fixed		A	$\mathrm{A8}_f$
INPO	log-ratio to reference/iterate mix	squared	equal	mixture		B_t	B_t
DNO	log-ratio to prev. iterate	soft-label logistic	pair	prev. iterate		C_t	C_t
TDPO	log-ratio plus sequence KL	logistic	equal	fixed		D	D
R-DPO	log-ratio plus length bonus	logistic	equal	fixed		E	E
α-DPO	length-norm. log-lik. minus data offset	logistic	pair	mixture		F	F
SimPO	length-normalised log-likelihood	shifted logistic	equal	none		G	G1
RRHF	length-normalised log-likelihood	hinge, m=0	equal	none	SFT	G	G2
ORPO	log-odds of length-norm. likelihood	logistic	equal	none	SFT	G	G3
CPO	log-likelihood	logistic	equal	none	NLL	H	H1
SLiC-HF	log-likelihood	hinge, m=δ	equal	none	CE, ref. text	H	H2
RankNet	generic scorer	soft-label logistic	equal	none		I	I1
LambdaRank	generic scorer	logistic	pair	none		I	I2
Per-response objectives (no pairwise margin)							
KTO	sigmoid of log-ratio minus batch ref. point	per-response	response	fixed		A	—
BCO	log-ratio minus running ref. point	per-response	equal	fixed		A	—
APO (anchored)	log-ratio to reference	per-response	equal	fixed		A	—
SPPO	log-ratio to prev. iterate	per-resp. squared	equal	prev. iterate		J_t	—
Policy-gradient and token-level objectives (no pairwise margin or per-response score)							
GRPO	group-normalised advantage		group	fixed			
RLOO	leave-one-out baseline advantage		group (cancels)	fixed			
ReMax	greedy-rollout baseline advantage		prompt	fixed			
PPO-RLHF	critic-based advantage			fixed	pretraining		
Nash-MD-PG	preference-oracle advantage		pair	mixture			
RTO	token-level reward from a DPO policy			fixed			
Listwise (no pairwise decomposition beyond two responses)							
ListNet	generic scorer	softmax CE		none		I	—
ListMLE	generic scorer	Plackett–Luce		none		I	—

Table 6: Experiment 1 at step 300, mean of three seeds (every standard deviation ≤ 0.026). "Exact" identities hold by construction.

Pair	Certificate	ρ	sign agr.	τ_{K}
DPO vs DPO, second formula (control)	Loss, exact	0.000	1.000	1.000
DPO vs random (control)	none	0.900	0.501	0.003
DPO vs IPO	Score, exact	0.000	1.000	1.000
DPO vs KTO score	Rank, exact; Score fails	0.075	1.000	1.000
SimPO vs ORPO	Rank, exact; Score fails	0.003	1.000	1.000
SLiC vs SimPO	none (length normalisation)	0.549	0.740	0.480
SLiC vs ORPO	none	0.547	0.740	0.480
DPO vs SimPO	none (reference)	0.778	0.593	0.187
DPO vs ORPO	none	0.776	0.593	0.187
DPO vs SLiC	none	0.697	0.630	0.263

Protocol. We train Qwen2.5-0.5B-Instruct with DPO on `ultrafeedback_binarized` ($\beta = 0.1$, learning rate 5×10^{-6}, batch 32, 300 steps, three seeds) and save checkpoints at steps 100, 200 and 300. The probe set has 3 000 candidates: six responses for each of 500 held-out prompts, four sampled from the reference model plus the dataset's chosen and rejected responses. At each checkpoint one forward pass gives every quantity that six methods' scores need. SLiC denotes SLiC-HF's score, the plain log-likelihood of the response, with no reference model and no length normalisation. Because every score comes from that one pass, the certified rows of Table 6 hold by construction and check

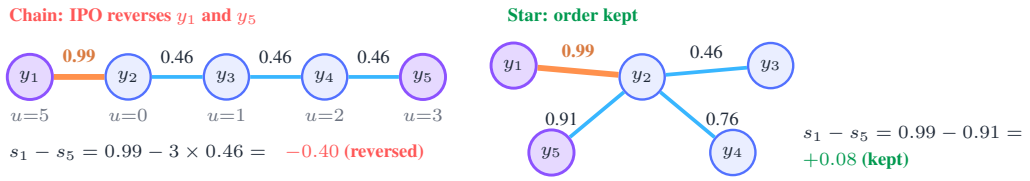

Figure 2: Why a chain can reverse a preference (IPO, Bradley–Terry labels, y_1 preferred 88/12). Edge labels are fitted margins. IPO's target map is bounded by $m = 1$, so it caps the large step at 0.99 and three small steps of 0.46 outweigh it along the chain. The star compares both ends with one anchor.

the implementation. The uncertified rows are the finding. A second run at learning rate 5×10^{-7} keeps the policy near the reference model (Appendix E).

Metrics. For each pair of methods we compare their scores over each prompt's six candidates. *Sign agreement* is the fraction of response pairs that both methods order the same way. It is 1 exactly when they share a ranking. The *affine residual* ρ is the fraction of variance left after the best per-prompt rescale and shift from one score to the other. It is 0 at a checkpoint exactly when the two scores are per-prompt affine images of each other there. Kendall's rank correlation τ_K sizes any disagreement. Two controls anchor the metrics. DPO against its own score computed by a second formula must give $\rho = 0$ and agreement 1. DPO against random scores must give agreement near 0.5.

Reading. The controls behave as specified at every checkpoint and seed. DPO/IPO give $\rho = 0$. The two pairs certified only at Rank give agreement and τ_K of one, with residuals of 0.003 (SimPO/ORPO) and 0.075 (DPO/KTO), so their scores agree on every winner without being rescalings of each other. Removing length normalisation (SLiC against SimPO or ORPO) lowers sign agreement to 0.74. Any reference-based score against any reference-free one agrees on 0.57 to 0.63 of pairs at every checkpoint, above the random control's 0.49 to 0.52, with τ_K of 0.19 to 0.26 at step 300. The agreement depends on how far the policy has moved from the reference. In the near-reference run it is 0.50 to 0.55, within noise of the random control (Table 12).

Experiment 2: which pairs get labelled, and the link

Two methods share a score and differ only in the link. Can the choice of which pairs to label change which response the trained policy prefers? Yes. The experiment is solved in closed form, so every number is reproducible to the digit.

Setup. Each response has a hidden utility. The win probability of a labelled pair is a fixed increasing function F of the utility gap, the *labelling function*. We use the two standard choices, Bradley–Terry (logistic, Bradley & Terry, 1952) and Thurstone (normal CDF, Thurstone, 1927). A *design* is the set of pairs that get labelled. A *chain* labels neighbours along some ordering of the responses. A *star* labels every response against one anchor. The *complete* design labels every pair. Scores are free parameters, so each loss places every labelled pair exactly at its target.

Result 2. *A star or complete design recovers the true order for every link of Section 4. A chain recovers it for all utilities only when the link's target map is proportional to the inverse of the labelling function. Under Bradley–Terry labels that is the logistic link alone.* The formal statement is Proposition 4, and Figure 2 shows the mechanism. Table 7 gives two exact instances. Under Bradley–Terry labels IPO and cDPO reverse an 88/12 preference that DPO keeps. Under Thurstone labels DPO's target map grows faster than the utility gap, so one large step outweighs four small ones, and DPO reverses a pair that IPO keeps. Whether a link is safe depends on the link and on how the labels were generated.

Sweep. Over 10 000 random utility draws with six responses, chains in random order reverse 9.2% of unlabelled pairs under IPO with Bradley–Terry labels and 4.8% under DPO with Thurstone labels. Stars, complete designs and chains in utility order reverse none. Adding the unlabelled pairs with their true labels at a reduced weight w helps only partly. At $w = 0.3$ IPO reverses 2.0% of pairs. Full weight, the complete design, brings every link below 1% (Appendix E.2).

Experiment 3: target maps in a trained model

Table 7: Experiment 2, two separate instances: fitted margin between the first and last response. "True gap" is their utility difference and p the win probability it implies. Negative entries are reversals.

		chain			star (anchor: worst response)		
Labels, utilities	true gap (p)	DPO	IPO	cDPO	DPO	IPO	cDPO
BT, $(5, 0, 1, 2, 3)$	$+2$ (0.88)	$+2.00$	$\mathbf{-0.40}$	$\mathbf{-0.19}$	$+2.00$	$+0.08$	$+0.31$
Thurstone, $(0, 4, 3, 2, 1, -1)$	$+1$ (0.84)	$\mathbf{-1.60}$	$+2.00$	$+3.49$	$+1.67$	$+0.68$	$+1.23$

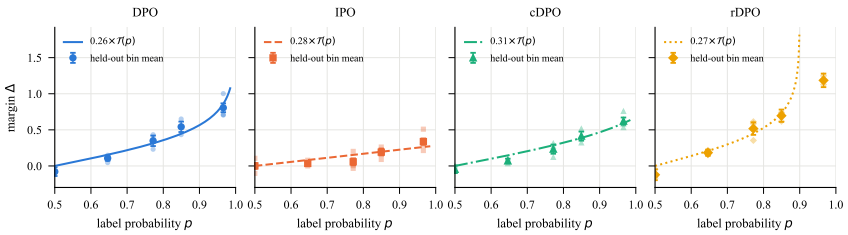

Figure 3: Experiment 3: held-out bin means of Δ against p at step 200, with each target map scaled by its fitted slope. Error bars are standard errors; small markers are single seeds.

Do four links that share a score train towards margins ordered as their target maps predict? Yes, in every seed. Result 1 describes the population minimiser. This experiment tests the part of it that a short training run can reach.

Setup. Same initialisation and data as Experiment 1, three seeds. Labels are soft. Each training pair carries $p = \sigma(\kappa(s_c - s_r))$, computed from the per-response ratings s_c, s_r that the dataset gives its chosen and rejected response, with κ fixed so that the mean of p is 0.75. DPO, IPO (target $m = 1$), cDPO ($\epsilon = 0.1$) and rDPO ($\epsilon = 0.1$) are trained on the same margin $\Delta = \beta[\log(\pi_\theta/\pi_{\text{ref}})(y) - \log(\pi_\theta/\pi_{\text{ref}})(y')]$ at $\beta = 0.1$ for 200 steps. On held-out pairs binned by p we record the mean margin per bin (Figure 3; Table 13 in Appendix E.1 gives the numbers).

Reading. In each of the three bins with $p \in (0.5, 0.9)$ the margins order rDPO $>$ DPO $>$ cDPO $>$ IPO, in each seed separately, and 8 of the 9 adjacent differences exceed two standard errors. No two curves are proportional. Every pair has a cross-bin ratio that varies by 17% or more. After 200 steps the margins reach 26% to 31% of their targets, and at that distance the curve shapes follow the target maps only in part. IPO's margins stay below its bound of 1, DPO's curve is convex, and rDPO's is not.

7 What this means in practice

- *Name the level before switching methods.* Moving from DPO to IPO, cDPO or rDPO keeps every ranking the model learns and changes the margin it aims for. Moving from DPO to SimPO changes the ranking itself.
- *The link sets how far training goes.* Under hard labels, a link with a bounded target map (IPO, cDPO) stops at a fixed margin. A link with an unbounded one (DPO, rDPO) keeps pushing the policy away from the reference model.
- *Label stars or all pairs.* A chain can reverse a preference whenever the link and the labelling function do not match, and down-weighting the unlabelled pairs leaves some reversals in place.
- *GRPO's normalisation is a per-group temperature.* Changing the group changes how hard a pair is pushed and leaves which response wins unchanged, so group size and composition belong next to β in a report.
- *State the level and give a witness.* Two methods that agree at Rank and differ at Loss should be reported as such, with the witness that separates them.

Limitations. The certificates are algebraic. They say nothing about finite samples, optimisers or learning rates, and Experiment 3 shows margins reaching about a quarter of their targets after 200 steps. The loss cards are hand transcriptions. A transcription error changes a row, which is why every card cites its source equation and the scripts are released. Experiments 1 and 3 use one 0.5B model and one dataset, and the gap between reference-based and reference-free scores is measured there,

without an account of how it scales. Experiment 2 treats scores as free parameters, so every labelled pair sits at its target. A real model shares parameters across prompts, which can damp or amplify the reversals.

8 Related work

Unifying templates. ΨPO (Azar et al., 2024), GPO (Tang et al., 2024b), RPO (Sun et al., 2025), RainbowPO (Zhao et al., 2025) and the unification of Raheja & Pochhi (2026) each write many objectives as instances of one template, and the surveys of Liu et al. (2025a) and Jiang et al. (2024) organise the same methods by design choice. None of these works tests one method against another, and none says at which level two instances of a template agree. The levels locate the components RainbowPO ablates. Length normalisation changes the ranking class, label smoothing changes the link under one score, and SFT mixing adds a term outside the margin.

Equivalence of reward functions. The closest formal precedent asks when two reward functions are interchangeable. Ng et al. (1999) show that potential-based shaping is the only additive change that preserves every optimal policy. EPIC (Gleave et al., 2021) and STARC (Skalse et al., 2024) canonicalise a reward to remove shaping, normalise its scale and compare what remains, and Skalse et al. (2023) characterise which transformations of a reward each common data source and downstream task cannot distinguish. Our per-prompt normalisation of scores (Section 3) is the one-step instance of this construction, applied to the score of a loss. Two things change when the object is a loss. The link and the weight become components that can differ under one score, and a per-prompt scale, which preserves the optimal action, changes the KL-regularised optimum (Proposition 3).

Surrogate losses. The target map is the link function of a proper composite loss in the sense of Reid & Williamson (2010), the map from class probability to the risk-minimising prediction, and our link is their composite loss read as a function of the prediction. Calibration is the pointwise Fisher consistency of Bartlett et al. (2006), and Duchi et al. (2010) show that pairwise surrogates can be inconsistent for ranking, the population analogue of the reversals in Experiment 2.

DPO, RLHF and optima. Xu et al. (2024b) prove that the policies PPO reaches are a strict subset of DPO's. Yang et al. (2026) give conditions under which DPO and RLHF coincide. Rafailov et al. (2024b) derive DPO inside the token-level MDP. Tang et al. (2024a) and Shi et al. (2025) analyse the gaps between online and offline training and between RLHF and DPO. These works compare objectives with no common loss, at the level of optima or trajectories. That is the reason Opt sits outside the chain. Appendix F covers pair construction, learning dynamics, GRPO and learning to rank.

9 Conclusion

A claim that two objectives are equivalent needs a level, and the released loss cards and scripts decide that level for any pair of the 33 methods. Three questions are open. First, Experiment 3 shows four objectives that share a score ordering their margins as their target maps predict, after 200 steps on a 0.5B model. Whether the margins reach their targets, and whether the ordering holds at scale, needs larger models and longer runs. Second, the target map reads a link as a choice of how far a labelled pair is pushed. A link can therefore be designed from the labelling model of the data, and Proposition 4 gives the condition it must satisfy on sparse designs. Third, policy-gradient objectives have no score, so the rank and score levels do not apply to them. GRPO's per-group temperature (Corollary 2) is one relationship among such objectives that the four levels do not name, and extending the levels to advantage-weighted objectives is open.

AI use statement

In this work we used generative AI tools for implementing methods (the verification scripts and the experiment code of Section 6 and Appendix E), for cleaning and reformatting data (extracting and transcribing objectives from source files into loss cards), and for interpreting results. We have not used generative AI tools to develop the theoretical framework, to formulate the mathematical claims, to provide ingredients for or write the proofs, to propose hypotheses, or to design the methodology or the experiments. Synthetic data generation, translation and qualitative analysis are not applicable

to this work. Additionally, we used generative AI tools for identifying literature, editing the paper for readability and formatting references. We have reviewed all AI-assisted work. Every extracted equation was checked against the source location it cites, every verification script was run and its printed output inspected, and the target-map table was recomputed independently after a discrepancy was found in a tool-generated version. We take responsibility for the final content of this work, including text, claims and artefacts produced with the aid of generative AI.

Ethics statement

This work analyses published objectives and runs a small experiment on a public model and dataset. It involves no human subjects. The only ethical exposure is the accuracy of the claims it makes about prior work, which are held to a source-level standard.

Reproducibility statement

Appendix B proves every stated result. Appendix C derives every target map. Appendix D gives, for each of the 33 methods, the equation and file location of the objective used and the derivation of its row in Table 5. Appendix E gives the full configuration of every training run and checkpoint. The supplementary material contains the code and results of all three experiments (the probe set, the scoring and metric code, the trainers, the per-checkpoint scores and margins, and one script that regenerates every number of Experiment 2 with a check that asserts Table 7), the loss cards, and verification scripts that recompute every closed-form number printed in the paper. Its README maps each folder to the section, table or figure it supports.

References

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In *Association for Computational Linguistics*, 2024.

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 *International Conference on Artificial Intelligence and Statistics*, pp. 4447–4455, 2024.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *International Conference on Machine Learning*, pp. 89–96, 2005.

Christopher J Burges, Robert Ragno, and Quoc V Le. Learning to rank with nonsmooth cost functions. In *Advances in Neural Information Processing Systems*, volume 19, 2006.

Christopher J C Burges. From RankNet to LambdaRank to LambdaMART: An overview. Technical Report MSR-TR-2010-82, Microsoft Research, 2010.

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

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: From pairwise approach to listwise approach. In *International Conference on Machine Learning*, pp. 129–136, 2007.

Angelica Chen, Sadhika Malladi, Lily H Zhang, Xinyi Chen, Qiuyi Zhang, Rajesh Ranganath, and Kyunghyun Cho. Preference learning algorithms do not learn preference rankings. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Yurong Chen, Yu He, Michael I. Jordan, and Fan Yao. How sampling shapes LLM alignment: From one-shot optima to iterative dynamics. *arXiv preprint arXiv:2602.12180*, 2026.

Sayak Ray Chowdhury, Anush Kini, and Nagarajan Natarajan. Provably robust DPO: Aligning language models with noisy feedback. In *International Conference on Machine Learning*, 2024.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Karel D'Oosterlinck, Winnie Xu, Chris Develder, Thomas Demeester, Amanpreet Singh, Christopher Potts, Douwe Kiela, and Shikib Mehri. Anchored preference optimization and contrastive revisions: Addressing underspecification in alignment. *arXiv preprint arXiv:2408.06266*, 2024.

John C Duchi, Lester W Mackey, and Michael I Jordan. On the consistency of ranking algorithms. In *International Conference on Machine Learning*, 2010.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. KTO: Model alignment as prospect theoretic optimization. In *International Conference on Machine Learning*, 2024.

Adam Fisch, Jacob Eisenstein, Vicky Zayats, Alekh Agarwal, Ahmad Beirami, Chirag Nagpal, Pete Shaw, and Jonathan Berant. Robust preference optimization through reward model distillation. *Transactions on Machine Learning Research*, 2025.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866, 2023.

Adam Gleave, Michael Dennis, Shane Legg, Stuart Russell, and Jan Leike. Quantifying differences in reward functions. In *International Conference on Learning Representations*, 2021.

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

Shawn Im and Yixuan Li. Understanding the learning dynamics of alignment with human feedback. In *International Conference on Machine Learning*, 2024.

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17176–17186, 2020.

Ruili Jiang, Kehai Chen, Xuefeng Bai, Zhixuan He, Juntao Li, Muyun Yang, Tiejun Zhao, Liqiang Nie, and Min Zhang. A survey on human preference learning for large language models. *arXiv preprint arXiv:2406.11191*, 2024.

Seungjae Jung, Gunsoo Han, Daniel Wontae Nam, and Kyoung-Woon On. Binary classifier optimization for large language model alignment. *arXiv preprint arXiv:2404.04656*, 2024.

Tomasz Korbak, Ethan Perez, and Christopher Buckley. RL with KL penalties is better viewed as Bayesian inference. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022.

Ziniu Li, Tian Xu, Yushun Zhang, Zhihang Lin, Yang Yu, Ruoyu Sun, and Zhi-Quan Luo. ReMax: A simple, effective, and efficient reinforcement learning method for aligning large language models. In *International Conference on Machine Learning*, 2024.

Ning Liu, Chuanneng Sun, Kristina Klinkner, and Shervin Malmasi. Beyond pairs: Your language model is secretly optimizing a preference graph. *arXiv preprint arXiv:2605.08037*, 2026.

Shunyu Liu, Wenkai Fang, Zetian Hu, Junjie Zhang, Yang Zhou, Kongcheng Zhang, Rongcheng Tu, Ting-En Lin, Fei Huang, Mingli Song, Yongbin Li, and Dacheng Tao. A survey of direct preference optimization. *arXiv preprint arXiv:2503.11701*, 2025a.

Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu Liu. Statistical rejection sampling improves preference optimization. In *International Conference on Learning Representations*, 2024.

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. *arXiv preprint arXiv:2503.20783*, 2025b.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.

Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Eric Mitchell. A note on DPO with noisy preferences and relationship to IPO. `https://ericmitchell.ai/cdpo.pdf`, 2023.

Youssef Mroueh. Reinforcement learning with verifiable rewards: GRPO's effective loss, dynamics, and success amplification. *arXiv preprint arXiv:2503.06639*, 2025.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. Nash learning from human feedback. In *International Conference on Machine Learning*, 2024.

Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *International Conference on Machine Learning*, pp. 278–287, 1999.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4998–5017, 2024.

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 *Advances in Neural Information Processing Systems*, volume 36, 2023.

Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sikchi, Joey Hejna, Bradley Knox, Chelsea Finn, and Scott Niekum. Scaling laws for reward model overoptimization in direct alignment algorithms. In *Advances in Neural Information Processing Systems*, volume 37, 2024a.

Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. From r to Q^*: Your language model is secretly a Q-function. *arXiv preprint arXiv:2404.12358*, 2024b.

Tarun Raheja and Nilay Pochhi. From RLHF to direct alignment: A theoretical unification of preference learning for large language models. *arXiv preprint arXiv:2601.06108*, 2026.

Arun Rajkumar and Shivani Agarwal. A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In *International Conference on Machine Learning*, 2014.

Noam Razin, Sadhika Malladi, Adithya Bhaskar, Danqi Chen, Sanjeev Arora, and Boris Hanin. Unintentional unalignment: Likelihood displacement in direct preference optimization. *arXiv preprint arXiv:2410.08847*, 2024.

Mark D Reid and Robert C Williamson. Composite binary losses. *Journal of Machine Learning Research*, 11:2387–2422, 2010.

Corby Rosset, Ching-An Cheng, Arindam Mitra, Michael Santacroce, Ahmed Awadallah, and Tengyang Xie. Direct Nash optimization: Teaching language models to self-improve with general preferences. *arXiv preprint arXiv:2404.03715*, 2024.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Nihar B. Shah and Martin J. Wainwright. Simple, robust and optimal ranking from pairwise comparisons. *Journal of Machine Learning Research*, 18(199):1–38, 2018.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Ruizhe Shi, Minhak Song, Runlong Zhou, Zihan Zhang, Maryam Fazel, and Simon S Du. Understanding the performance gap in preference learning: A dichotomy of RLHF and DPO. *arXiv preprint arXiv:2505.19770*, 2025.

Joar Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, and Adam Gleave. Invariance in policy optimisation and partial identifiability in reward learning. In *International Conference on Machine Learning*, 2023.

Joar Skalse, Lucy Farnik, Sumeet Ramesh Motwani, Erik Jenner, Adam Gleave, and Alessandro Abate. STARC: A general framework for quantifying differences between reward functions. In *International Conference on Learning Representations*, 2024.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

Shengyang Sun, Yian Zhang, Alexander Bukharin, David Mosallanezhad, Jiaqi Zeng, Soumye Singhal, Gerald Shen, Adithya Renduchintala, Tugrul Konuk, Yi Dong, Zhilin Wang, Dmitry Chichkov, Olivier Delalleau, and Oleksii Kuchaiev. Reward-aware preference optimization: A unified mathematical framework for model alignment. *arXiv preprint arXiv:2502.00203*, 2025.

Fahim Tajwar, Anikait Singh, Archit Sharma, Rafael Rafailov, Jeff Schneider, Tengyang Xie, Stefano Ermon, Chelsea Finn, and Aviral Kumar. Preference fine-tuning of LLMs should leverage suboptimal, on-policy data. In *International Conference on Machine Learning*, 2024.

Yunhao Tang, Daniel Zhaohan Guo, Zeyu Zheng, Daniele Calandriello, Yuan Cao, Eugene Tarassov, Rémi Munos, Bernardo Avila Pires, Michal Valko, Yong Cheng, and Will Dabney. Understanding the performance gap between online and offline alignment algorithms. *arXiv preprint arXiv:2405.08448*, 2024a.

Yunhao Tang, Zhaohan Daniel Guo, Zeyu Zheng, Daniele Calandriello, Rémi Munos, Mark Rowland, Pierre Harvey Richemond, Michal Valko, Bernardo Avila Pires, and Bilal Piot. Generalized preference optimization: A unified approach to offline alignment. In *International Conference on Machine Learning*, 2024b.

Louis L. Thurstone. A law of comparative judgment. *Psychological Review*, 34(4):273–286, 1927.

Milan Vojnovic and Se-Young Yun. What is the alignment objective of GRPO? *arXiv preprint arXiv:2502.18548*, 2025.

Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. Beyond reverse KL: Generalizing direct preference optimization with diverse divergence constraints. *arXiv preprint arXiv:2309.16240*, 2023.

Junkang Wu, Xue Wang, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. α-DPO: Adaptive reward margin is what direct preference optimization needs. *arXiv preprint arXiv:2410.10148*, 2024a.

Yihong Wu, Liheng Ma, Lei Ding, Muzhi Li, Xinyu Wang, Kejia Chen, Zhan Su, Zhanguang Zhang, Chenyang Huang, Yingxue Zhang, Mark Coates, and Jian-Yun Nie. It takes two: Your GRPO is secretly DPO. *arXiv preprint arXiv:2510.00977*, 2025.

Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. Self-play preference optimization for language model alignment. *arXiv preprint arXiv:2405.00675*, 2024b.

Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise approach to learning to rank: Theory and algorithm. In *International Conference on Machine Learning*, pp. 1192–1199, 2008.

Tengyang Xie, Dylan J Foster, Akshay Krishnamurthy, Corby Rosset, Ahmed Awadallah, and Alexander Rakhlin. Exploratory preference optimization: Harnessing implicit Q*-approximation for sample-efficient RLHF. *arXiv preprint arXiv:2405.21046*, 2024.

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 *International Conference on Machine Learning*, 2024a.

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

Zhiqin Yang, Yonggang Zhang, Wei Xue, Dong Fang, Bo Han, and Yike Guo. Conditional equivalence of DPO and RLHF: Implicit assumption, failure modes, and provable alignment. *arXiv preprint arXiv:2605.20834*, 2026.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. RRHF: Rank responses to align language models with human feedback without tears. In *Advances in Neural Information Processing Systems*, 2023.

Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. Token-level direct preference optimization. In *International Conference on Machine Learning*, 2024.

Yuheng Zhang, Dian Yu, Baolin Peng, Linfeng Song, Ye Tian, Mingyue Huo, Nan Jiang, Haitao Mi, and Dong Yu. Iterative Nash policy optimization: Aligning LLMs with general preferences via no-regret learning. *arXiv preprint arXiv:2407.00617*, 2024.

Hanyang Zhao, Genta Indra Winata, Anirban Das, Shi-Xiong Zhang, David D Yao, Wenpin Tang, and Sambit Sahu. RainbowPO: A unified framework for combining improvements in preference optimization. In *International Conference on Learning Representations*, 2025.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. SLiC-HF: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*, 2023.

Han Zhong, Zikang Shan, Guhao Feng, Wei Xiong, Xinle Cheng, Li Zhao, Di He, Jiang Bian, and Liwei Wang. DPO meets PPO: Reinforced token optimization for RLHF. In *International Conference on Machine Learning*, 2025.

Wenxuan Zhou, Ravi Agrawal, Shujian Zhang, Sathish Reddy Indurthi, Sanqiang Zhao, Kaiqiang Song, Silei Xu, and Chenguang Zhu. WPO: Enhancing RLHF with weighted preference optimization. In *Empirical Methods in Natural Language Processing*, 2024.

Guide to the appendix. A states every result formally. B gives the proofs and the witness types. C derives each target map. D gives each method's loss card with its source location. E gives the experimental configurations and the details of Experiment 3. F covers further related work.

A Formal statements

This appendix states formally every result that the main text gives in words. Proofs follow in Appendix B.

A.1 LEVELS

Two objectives with scores s_1, s_2 on the same policy class are compared. Write Δ_1, Δ_2 for their margins.

Definition 1 (Levels of equivalence)**.**
- Rank***, same ordering.*** *At every parameter value, on every prompt, the two methods agree on which of any two responses scores higher:* $\operatorname{sign}\Delta_1 = \operatorname{sign}\Delta_2$ *for all* θ*,* x *and* $y, y' \in \mathcal{Y}_x$*. Equivalently,* $s_1(x,\cdot) = g_{x,\theta} \circ s_2(x,\cdot)$ *for some strictly increasing* $g_{x,\theta}$*.*
- Score***, same score.*** *One score is a positive rescaling and shift of the other, where the rescaling may depend on the prompt and not on* θ*:* $s_1(x,y) = a(x)\, s_2(x,y) + b(x)$ *with* $a(x) > 0$*.*
- Loss***, same loss.*** Score *holds, the links agree once the rescaling is absorbed,* $\ell_1(a(x)u) = c\,\ell_2(u) + d$ *for constants* $c > 0$*,* d*, the weights are proportional,* $\omega_1 = \kappa\,\omega_2$ *with a constant* $\kappa > 0$*, and neither objective has extra terms. Then* $\mathcal{L}_1 = c\kappa\,\mathcal{L}_2 + \text{const}$ *as functions of* θ*, on any* $\mathcal{D}$*.*
- Opt***, same optimum*** *on a data distribution* $\mathcal{D}$*: the population minimisers over the policy class coincide.*

Rank and Score are properties of the score alone, so they apply to per-response objectives too. Opt is a statement about one particular $\mathcal{D}$.

Proposition 1 (Implications and separations)**.** Loss $\Rightarrow$ Score $\Rightarrow$ Rank*, and* Loss *together with a common* $\mathcal{D}$ *implies* Opt*. No implication reverses, and* Opt *implies none of the others: (a)* Score $\not\Rightarrow$ Loss*: DPO and IPO; (b)* Rank $\not\Rightarrow$ Score*: SimPO and ORPO; (c)* Score $\not\Rightarrow$ Opt*, hence* Rank $\not\Rightarrow$ Opt*: a score and the same score multiplied by a positive per-prompt factor; (d)* Opt $\not\Rightarrow$ Loss*, also under a shared score: two links with proportional target maps (Theorem 1), or the pessimistic-DPO/IPO pair of Fisch et al. (2025), which share an unconstrained optimum with different loss surfaces; (e)* Opt $\not\Rightarrow$ Rank*, hence* Opt $\not\Rightarrow$ Score*:* Opt *concerns one* $\mathcal{D}$ *and* Rank *quantifies over every prompt and* θ*, so two objectives that coincide on the support of* $\mathcal{D}$ *and order some pair differently on a prompt outside it share the minimiser and fail* Rank*.*

(a) and (b) are the pairs of Section 3 made exact, and (c) is Proposition 3.

Deciding a level. Given two cards, Score and Loss are decided mechanically by a *canonical form*: at Score the score centred and scale-fixed per prompt, $\Phi(x,y) = (s(x,y) - \bar{s}(x))/\operatorname{med}_y |s(x,y) - \bar{s}(x)|$, which a per-prompt affine map leaves unchanged. Equal canonical forms give $s_1 = a(x,\theta)\, s_2 + b(x,\theta)$, and Definition 1 asks for a scale free of θ, so the ratio of the two spreads is read off the formulas as well and must not depend on θ (for DPO and IPO it is the constant β); two parameter settings at which the ratio differs witness a failure. At Loss the same with the link normalised to $\ell(0) = 0$, $\ell'(0) = -1$, the weight divided by its mean, and $\mathcal{E} = \varnothing$. Rank is certified by exhibiting the strictly increasing map g of Definition 1 (for ORPO against SimPO, $g = \operatorname{logit} \circ \exp(\cdot/\beta)$) and refuted by two pairs with margins of opposite sign, which a numerical search over responses supplies. Opt is certified only by a theorem, here Theorem 1(ii) under its realisability hypothesis, and refuted by Proposition 3(b).

A.2 TARGET MAPS AND SHARED SCORES

Definition 2 (Target map)**.** *Fix a link* ℓ *and a pair whose label says the first response wins with probability* $p \in (0,1)$*. The* target map $\mathcal{T}_\ell(p)$ *is the margin that minimises the pair's expected loss,* $\mathcal{T}_\ell(p) = \arg\min_{\Delta \in \overline{\mathbb{R}}}\; p\,\ell(\Delta) + (1-p)\,\ell(-\Delta)$*, taken as a set when the minimiser is not unique. A link is* calibrated *if* $\mathcal{T}_\ell$ *is single-valued, strictly increasing on* $(0,1)$ *and* $\mathcal{T}_\ell(\frac{1}{2}) = 0$*.*

Calibration is the pointwise Fisher consistency of Bartlett et al. (2006) for pair classification.

Proposition 2 (Closed forms)**.** *Table 8 gives* $\mathcal{T}_\ell$ *for every link in use (derivations in Appendix C). All except the hinge and the DNO link are calibrated, rDPO on* $(\epsilon, 1-\epsilon)$*, where its map is finite. The cDPO and rDPO maps are the logistic map at a re-noised and a de-noised label (Appendix C).*

Theorem 1 (What a shared score certifies)**.** *Let two margin objectives be* Score*-equivalent with* $s_1 = a(x)s_2 + b(x)$*, weights depending on* x *only, no extra terms, and calibrated links* ℓ_1, ℓ_2*. Then (i) they are* Rank*-equivalent at every* θ*; (ii) suppose the score class can realise any transitive margin (every assignment* $\Delta(y,y') = v(y) - v(y')$ *on the support of* $\mathcal{D}$*) and the targets of both links are jointly realisable: for each prompt there is a* v *with* $v(y) - v(y') = \mathcal{T}_\ell(p_{yy'})$ *for every labelled pair,*

Table 8: Target maps with link formulas. m, ϵ, γ, z are the link's own parameters; $\operatorname{logit} p = \log \frac{p}{1-p}$.

Link $\ell(\Delta)$	$\mathcal{T}_\ell(p)$	Used by
$-\log \sigma(\Delta)$	$\operatorname{logit} p$ (unbounded)	DPO, ORPO, CPO, R-DPO, WPO, RSO, f-DPO, LambdaRank
$(\Delta - m)^2$	$m(2p-1)$ (bounded by $\pm m$)	IPO ($m = \frac{1}{2\tau}$), IPO-MD, INPO
$\max(0, m - \Delta)$	$m \operatorname{sign}(2p-1)$; any point of $[-m, m]$ at $p = \frac{1}{2}$	SLiC-HF, RSO-hinge, RRHF ($m = 0$)
$-(1-\epsilon)\log\sigma(\Delta) - \epsilon \log\sigma(-\Delta)$	$\operatorname{logit}(\epsilon + (1-2\epsilon)p)$, bounded by $\pm\operatorname{logit}(1-\epsilon)$	cDPO, RankNet (soft $\bar{P}$)
$\frac{-(1-\epsilon)\log\sigma(\Delta) + \epsilon\log\sigma(-\Delta)}{1-2\epsilon}$	$\operatorname{logit}\frac{p-\epsilon}{1-2\epsilon}$ on $(\epsilon, 1-\epsilon)$; $\pm\infty$ outside	rDPO
$-\log\sigma(\Delta - \gamma)$	root of $p\,\sigma(\gamma - \Delta) = (1-p)\,\sigma(\Delta + \gamma)$	SimPO, α-DPO
$-z\log\sigma(\Delta) - (1-z)\log\sigma(-\Delta)$	$\operatorname{logit} z$ at $p = 1$; depends on the pair	DNO (per-pair z)
$e^{-\Delta}$	$\frac{1}{2}\operatorname{logit} p$	(reference link)

which holds whenever the labelled pairs of a prompt form a forest (a chain or a star, with no cycle). Then each population minimiser places every labelled pair at its target, orders every labelled pair as $\operatorname{sign}(p - \frac{1}{2})$*, and the two minimisers coincide iff* $\mathcal{T}_{\ell_1}(p) = a(x)\,\mathcal{T}_{\ell_2}(p)$ *for all* p *in the support of* $\mathcal{D}$*. When the labelled pairs close a cycle the targets need not be jointly realisable (three responses with Bradley–Terry gaps* $1, 1, 2$ *give IPO targets* 0.462*,* 0.462 *and* 0.762*, and the first two force the third margin to* 0.924*), the minimiser is a compromise, and Proposition 4(c) treats the complete design; (iii) in general their gradients at a common* θ *need not share a sign, and their stationary points need not coincide.*

Corollary 1 (The DPO-score family)**.** *DPO, IPO, cDPO and rDPO have the same* Score *score* $\beta\log(\pi_\theta/\pi_{\text{ref}})$ *and four pairwise non-proportional target maps (*$\operatorname{logit} p$*,* $(2p-1)/2\tau$*,* $\operatorname{logit}(\epsilon + (1-2\epsilon)p)$*,* $\operatorname{logit}((p-\epsilon)/(1-2\epsilon))$*). On a forest design, the four population optima are pairwise distinct whenever the support of the label distribution contains an open interval of* $(0, 1)$ *(of* $(\epsilon, 1-\epsilon)$ *for rDPO). For a finitely supported label distribution they are distinct unless every label lies at a point where two of the maps agree, a set of isolated points that contains* $p = \frac{1}{2}$*, where every map is zero. Under hard labels DPO and rDPO have no finite minimiser, IPO's margins go to* $\pm 1/2\tau$ *and cDPO's to* $\pm\operatorname{logit}(1-\epsilon)$*.*

Per-response objectives. KTO's and BCO's loss on one response is monotone in its reward, so each reward is pushed to $\pm\infty$ relative to the reference point, as the logistic link pushes a margin under a hard label. SPPO's squared loss pushes each log-ratio to the finite target $\eta(\hat{P}(y \succ \pi_t) - \frac{1}{2})$, linear in the win probability, the form of IPO's pairwise target $(2p-1)/(2\tau)$.

Proposition 3 (Shift and scale)**.** *For the KL-regularised optimum* $\pi^* \propto \pi_{\text{ref}}\exp(r/\beta)$*: (a)* $r \mapsto r + f(x)$ *leaves* π^* *and the DPO loss unchanged; (b)* $r \mapsto r/\nu(x)$ *with* $\nu(x) > 0$ *preserves every ordering and changes* π^**, being the same as* $\beta \mapsto \beta\,\nu(x)$*.*

Corollary 2 (GRPO's normalisation)**.** *For a group* G *with rewards* r*, mean* μ_G *and standard deviation* sd_G*, GRPO's advantage* $A(y \mid G) = (r(y) - \mu_G)/\operatorname{sd}_G$ *satisfies* $A(y_1 \mid G) - A(y_2 \mid G) = (r(y_1) - r(y_2))/\operatorname{sd}_G$ *for every pair in* G*. The mean cancels; the standard deviation is a per-group instance of Proposition 3(b). Between two groups containing the same pair, the ratio of the effective margins is the single constant* $\operatorname{sd}_{G_2}/\operatorname{sd}_{G_1}$ *for every pair, so batch composition rescales and never reorders.*

A.3 Pair construction

Responses $y_1, \ldots, y_K$ have utilities $u_1, \ldots, u_K$ and labels $p_{ij} = F(u_i - u_j)$ for a strictly increasing, symmetric F (Bradley–Terry $F = \sigma$ (Bradley & Terry, 1952); Thurstone $F = \Phi$, the normal CDF (Thurstone, 1927)). A *design* is the set of labelled pairs: a *chain* labels consecutive pairs along an ordering of the responses, a *star* labels every response against one anchor, the *complete* design labels all pairs. Scores are unconstrained, the links of Corollary 1 act on the same margin, and $g = \mathcal{T}_\ell \circ F$ is the margin a link fits to a labelled pair with utility gap d.

Proposition 4 (Pair construction and the link)**.** *For a calibrated link with* $\mathcal{T}(1-p) = -\mathcal{T}(p)$*: (a)* Star*. The minimiser orders the responses as* u *does, for every such link and every* F*. (b)* Chain*. The minimiser sets every labelled pair's margin to its target, and the implied margin of an unlabelled pair is the sum of* g *over the path between them. This ordering agrees with* u *for all utilities and all chain lengths iff* g *is linear, i.e.* $\mathcal{T} = c\,F^{-1}$ *for some* $c > 0$*: under Bradley–Terry labels the logistic link alone, and under any other* F *the logistic link reverses some pair. (c)* Complete*. With equal weights,*

the logistic, squared and cDPO links (and rDPO when every $p_{ij} \in (\epsilon, 1-\epsilon)$) order the responses by the Borda score $\sum_j p_{ij}$, hence as u does, for every F.

(c) is the population form of Borda-count results (Shah & Wainwright, 2018).

B PROOFS

Proof of Proposition 1. Implications. Loss includes Score by definition. If $s_1 = a(x)s_2 + b(x)$ with $a > 0$ then $\Delta_1 = a(x)\Delta_2$ and the signs agree, so Score $\Rightarrow$ Rank. If two objectives are Loss-equivalent then $\mathcal{L}_1(\theta) = c\kappa\,\mathcal{L}_2(\theta) + \text{const}$ with $c\kappa > 0$, so the minimisers over the same policy class coincide on any common $\mathcal{D}$.

(a) DPO and IPO: $s_{\text{DPO}} = \beta s_{\text{IPO}}$, links logistic and squared. Their target maps are $\operatorname{logit} p$ and $(2p-1)/2\tau$, not related by a positive affine map in the argument, so the links are not related as Loss requires.

(b) SimPO and ORPO: with $P(y) = \exp(|y|^{-1} \log \pi_\theta(y \mid x)) \in (0,1)$, $s_{\text{SimPO}} = \beta \log P$ and $s_{\text{ORPO}} = \operatorname{logit} P$, both strictly increasing in P, so Rank holds at every θ. For Score one needs $\operatorname{logit} P = a\beta \log P + b$ on the range of P, which fails because logit and log have different curvature; the witness in Section 3 exhibits two pairs with equal Δ_{SimPO} and unequal Δ_{ORPO}, which no function of Δ_{SimPO}, affine or otherwise, can produce.

(c) Proposition 3(b).

(d) Theorem 1(ii) with $\mathcal{T}_{\ell_1} = a\mathcal{T}_{\ell_2}$ and $\ell_1 \neq c\,\ell_2(\cdot/a) + d$, for instance the logistic and exponential links with $a = \frac{1}{2}$ on a forest design under a score class that realises any transitive margin, or the pair in Fisch et al. (2025, Prop. 1).

(e) Let the two objectives have the same link and weight, and scores s_1, s_2 with $s_1(x,\cdot) = s_2(x,\cdot)$ for every x in the support of $\mathcal{D}$ and $s_1(x_0,\cdot) = -s_2(x_0,\cdot)$ on a prompt x_0 outside it. The two population losses are the same function of θ, so the minimisers coincide; at x_0 every non-zero margin has opposite signs under the two scores, so Rank fails. □

Proof of Theorem 1. (i) is the implication Score $\Rightarrow$ Rank. (ii) The population risk is a sum over labelled pairs of terms $p\,\ell(\Delta) + (1-p)\,\ell(-\Delta)$, each minimised at $\Delta = \mathcal{T}_\ell(p)$ alone, since the link is calibrated. When one v places every labelled pair at its target it attains the sum of the pairwise minima, so it is a minimiser and every minimiser has the same margins. On a forest such a v exists: fix one response per component and set the rest by walking the edges. Calibration gives $\operatorname{sign} \mathcal{T}_\ell(p) = \operatorname{sign}(p - \frac{1}{2})$ for both links. With $\Delta_1 = a(x)\Delta_2$, the first objective's minimiser expressed in Δ_2 is $\mathcal{T}_{\ell_1}(p)/a(x)$ and the second's is $\mathcal{T}_{\ell_2}(p)$; they coincide for all p in the support iff $\mathcal{T}_{\ell_1}(p) = a(x)\mathcal{T}_{\ell_2}(p)$ there. (iii) Proposition 5(c) gives gradients of opposite sign at one θ; for stationary points, take the two-pair soft-label population $\Delta_1 = \theta$ with $p_1 = 0.7$ and $\Delta_2 = -\frac{1}{2}\theta$ with $p_2 = 0.55$, for which the logistic risk is minimised at $\theta^* = 0.573$ and the exponential risk at $\theta^* = 0.290$ (grid search over $[-6, 6]$, step 10^{-5}). □

Proof of Corollary 1. The four scores are $\beta \log(\pi_\theta/\pi_{\text{ref}})$ up to the scalar β (IPO uses $\beta = 1$). Proposition 2 gives the four target maps. No two are proportional on an interval, since $\operatorname{logit} p$ is unbounded at both ends, $(2p-1)/2\tau$ is linear, $\operatorname{logit}(\epsilon + (1-2\epsilon)p)$ is bounded, and $\operatorname{logit}((p-\epsilon)/(1-2\epsilon))$ diverges at $p = \epsilon$ and $p = 1-\epsilon$, not at 0 and 1. For any two of the maps, $\mathcal{T}_{\ell_1} - a\,\mathcal{T}_{\ell_2}$ with a the scale between the two scores is a real-analytic function on $(0,1)$ (on $(\epsilon, 1-\epsilon)$ for rDPO) that is not identically zero, so its zeros are isolated, and $p = \frac{1}{2}$ is one of them. Theorem 1(ii) then separates the optima whenever the support meets the complement of that zero set, which every support containing an interval does. The hard-label statements are the limits $p \to 1$. □

Proposition 5 (Strict monotonicity is the wrong condition)**.** *(a) The squared link and the cDPO link are not monotone on $\mathbb{R}$ and are calibrated; monotonicity is not necessary for the ordering of optima to transfer. (b) The logistic and exponential links are both strictly decreasing and have target maps $\operatorname{logit} p$ and $\frac{1}{2}\operatorname{logit} p$, so with a common score they have different minimisers; monotonicity is not sufficient. (c) Two strictly decreasing, calibrated links can move a shared parameter in opposite directions at the same point. On a two-pair example the logistic gradient is -0.0258 and the exponential gradient $+0.1120$.*

Proof of Proposition 5. (a) The squared link $(\Delta - m)^2$ decreases on $(-\infty, m)$ and increases after; its target map $m(2p-1)$ is single-valued, increasing and zero at $\frac{1}{2}$. The cDPO link has derivative $\sigma(\Delta) - (1-\epsilon)$, which changes sign at $\Delta = \operatorname{logit}(1-\epsilon)$; its target map $\operatorname{logit}(\epsilon + (1-2\epsilon)p)$ is calibrated. (b) Both links have negative derivative everywhere; the target maps are $\operatorname{logit} p$ and $\frac{1}{2}\operatorname{logit} p$ (Appendix C). (c) Two pairs share one parameter θ, with $\Delta_1 = 2$, $\partial_\theta \Delta_1 = 1$, $\Delta_2 = -\frac{1}{2}$, $\partial_\theta \Delta_2 = -0.15$. Logistic: pair 1 contributes $-\sigma(-2) \cdot 1 = -0.1192$ and pair 2 contributes $-\sigma(0.5) \cdot (-0.15) = +0.0934$, total -0.0258. Exponential $e^{-\Delta}$: pair 1 contributes $-e^{-2} \cdot 1 = -0.1353$ and pair 2 contributes $-e^{0.5} \cdot (-0.15) = +0.2473$, total $+0.1120$. □

Proof of Proposition 3 and Corollary 2. (a) $\pi_{\text{ref}}(y) e^{(r(y)+f)/\beta} = e^{f/\beta} \pi_{\text{ref}}(y) e^{r(y)/\beta}$ and the factor cancels in the normaliser; the DPO margin is a difference of rewards at one x and f cancels. This is potential-based shaping with a one-step horizon, and Ng et al. (1999) prove the converse: a per-prompt shift is the *only* additive change that leaves every optimal policy unchanged, which is why the Score canonical form centres per prompt. (b) $\pi_{\text{ref}}(y) e^{r(y)/(\nu\beta)}$ is the optimum at temperature $\nu\beta$; orderings are preserved since $\nu > 0$; for $r = (1, 0.5, 0)$, uniform π_{ref}, $\beta = 1$ the optimum is $(0.506, 0.307, 0.186)$ at $\nu = 1$ and $(0.419, 0.326, 0.254)$ at $\nu = 2$. For the corollary, $A(y_1 \mid G) - A(y_2 \mid G) = ((r(y_1) - \mu_G) - (r(y_2) - \mu_G))/\mathrm{sd}_G$; the ratio across groups is $\mathrm{sd}_{G_2}/\mathrm{sd}_{G_1}$ for every pair since it does not involve the pair. RLOO's weight on y_i is $r_i - \frac{1}{k-1}\sum_{j \neq i} r_j$ with k the group size, so the difference of two weights is $(r_1 - r_2)(1 + \frac{1}{k-1})$, a scale that depends on k only; ReMax's is $r_i - r(\bar{y})$ with $\bar{y}$ the greedy rollout, and the difference is $r_1 - r_2$. □

Proposition 6 (SPPO is not a margin objective)**.** *SPPO's training loss (Wu et al., 2024b, eq. 11) is* $\mathbb{E}_{y \sim \pi_t}\big[(u(y) - c(y))^2\big]$ *with* $u(y) = \log(\pi_\theta(y \mid x)/\pi_t(y \mid x))$ *and* $c(y) = \eta(\hat{P}(y \succ \pi_t \mid x) - \frac{1}{2})$. *For a pair the two terms sum to* $\frac{1}{2}[(\Delta_u - \Delta_c)^2 + (\Sigma_u - \Sigma_c)^2]$ *with* $\Sigma_u = u(y) + u(y')$. *With* $\hat{P} = (1, 0)$ *and* $\eta = 1$*, so that* $c = (\frac{1}{2}, -\frac{1}{2})$*, the states* $(u, u') = (1, 0)$ *and* $(3, 2)$ *have the same margin, losses* 0.5 *and* 12.5*, and gradients* $(1, 1)$ *and* $(5, 5)$.

Proof. With $u = (\Sigma_u + \Delta_u)/2$, $u' = (\Sigma_u - \Delta_u)/2$ and likewise for c, $(u-c)^2 + (u'-c')^2 = \frac{1}{2}[(\Sigma_u - \Sigma_c)^2 + (\Delta_u - \Delta_c)^2]$. The stated numbers follow by substitution. □

Proof of Proposition 4. Symmetry of F and of $\mathcal{T}$ makes $g = \mathcal{T} \circ F$ odd, increasing, with $g(0) = 0$. On any tree design the objective separates over labelled pairs once the scores are parametrised by the differences along the edges, so each labelled pair's margin is set to its target $g(u_i - u_j)$. (a) With anchor y_0, $s_i - s_0 = g(u_i - u_0)$ is increasing in u_i. (b) On a path the implied margin between two responses is the sum of g over the steps between them. If $g(d) = cd$ the sum is $c(u_i - u_j)$. If g is not linear, $g(d)/d$ takes two different values at some $a, b > 0$; if $g(a)/a < g(b)/b$, a path with n_1 steps of $+a$ and n_2 steps of $-b$ has true gap $n_1 a - n_2 b > 0$ and fitted margin $n_1 g(a) - n_2 g(b) < 0$ whenever n_2/n_1 lies in the non-empty interval $(a/b,\ g(a)/g(b))$; if $g(a)/a > g(b)/b$, exchange the signs of the steps. Under Bradley–Terry labels $g = \mathcal{T} \circ \sigma$ is linear iff $\mathcal{T} = c \operatorname{logit}$, the logistic link; for any other F, $\operatorname{logit} \circ F$ is not linear. (c) For the squared link with target m the objective is $\sum_{i<j}[\Delta_{ij}^2 - 2m(2p_{ij} - 1)\Delta_{ij}]$ up to a constant, convex, with stationarity $K s_i - \sum_j s_j = m \sum_{j \neq i}(2p_{ij} - 1)$, so under the gauge $\sum_j s_j = 0$ the score s_i is an increasing affine function of $B_i = \sum_{j \neq i} p_{ij}$. For the logistic link stationarity reads $\sum_{j \neq i} \sigma(s_i - s_j) = B_i$; with $H(s) = \sum_j \sigma(s - s_j)$, which is strictly increasing, $H(s_i) = B_i + \frac{1}{2}$, so $s_i > s_k$ iff $B_i > B_k$. The cDPO loss at label p is the logistic loss at label $\epsilon + (1-2\epsilon)p$, and the rDPO loss at $p \in (\epsilon, 1-\epsilon)$ is the logistic loss at $(p - \epsilon)/(1 - 2\epsilon)$, both increasing in p, so the same ordering holds. Finally $B_i - B_k = \sum_{j \neq i,k}[F(u_i - u_j) - F(u_k - u_j)] + F(u_i - u_k) - F(u_k - u_i) > 0$ whenever $u_i > u_k$, for every increasing F. □

B.1 Witness types

B.2 GRPO worked witness

Worked witness: with responses $y_1, \ldots, y_4$ of rewards $(1.0, 0.6, 0.3, -0.2)$, $G_1 = \{y_1, y_2\}$ and $G_2 = \{y_1, \ldots, y_4\}$, the effective margins of (y_1, y_2) are 2.000 and 0.913 (population standard deviation, $\mathrm{sd}_{G_2} = 0.438$), a ratio equal to $\mathrm{sd}_{G_2}/\mathrm{sd}_{G_1} = 2.190$. The DeepSeekMath paper writes $\mathrm{std}(\mathbf{r})$ without saying whether it is the population or the sample deviation; the TRL reference

Table 9: Witness types: finite objects checkable without trusting the procedure that produced them.

Fails	Witness	Form
Score (transitivity)	cycle	y_1, y_2, y_3 with $\Delta(y_1, y_2) + \Delta(y_2, y_3) + \Delta(y_3, y_1) \neq 0$
Score (affinity)	non-affine score	pairs with Δ_2 equal and Δ_1 different at one x
Score (scale)	θ-dependent scale	θ, θ' at one x with different spread ratios
Loss (link)	target map	p with $\mathcal{T}_{\ell_1}(p) \neq a\,\mathcal{T}_{\ell_2}(p)$, a the score scale
Loss (weight)	pair weight	$(y_1, y_2), (y_3, y_4)$ at one x with $\omega(y_1, y_2) \neq \omega(y_3, y_4)$
Loss (weight)	batch weight	(y_1, y_2) in $G_1 \neq G_2$ with different effective margins
margin form	level	two states with equal Δ and different loss or gradient
margin form	extra term	a term in $\mathcal{E}$ evaluated on a response outside every pair

implementation (v1.13.0) uses the sample deviation with Bessel's correction, under which the worked witness of Section 4 gives effective margins 1.414 and 0.791 in place of 2.000 and 0.913.

C TARGET-MAP DERIVATIONS

Write $R(\Delta) = p\,\ell(\Delta) + (1-p)\,\ell(-\Delta)$.

Logistic. $R' = -p\,\sigma(-\Delta) + (1-p)\,\sigma(\Delta) = \sigma(\Delta) - p$, zero at $\operatorname{logit} p$.

Squared with target m. $R = p(\Delta - m)^2 + (1-p)(\Delta + m)^2$, $R' = 2\Delta - 2m(2p-1)$, minimiser $m(2p-1)$.

Hinge $\max(0, m - \Delta)$. R is piecewise linear with slope $-p + (1-p) = 1 - 2p$ on $(-m, m)$ and slope $-p$ below $-m$, $(1-p)$ above m; for $p > \frac{1}{2}$ it decreases up to $\Delta = m$ and increases after, so its unique minimiser is m; symmetrically $-m$ for $p < \frac{1}{2}$, and at $p = \frac{1}{2}$ every point of $[-m, m]$ minimises it. The map is constant on each side of $\frac{1}{2}$, so the hinge is not calibrated, and with $m = 0$ (RRHF) the target is 0 for every p.

cDPO. $\ell_\epsilon(\Delta) = -(1-\epsilon)\log\sigma(\Delta) - \epsilon\log\sigma(-\Delta)$. Collecting terms, $R = -q\log\sigma(\Delta) - (1-q)\log\sigma(-\Delta)$ with $q = p(1-\epsilon) + (1-p)\epsilon = \epsilon + (1-2\epsilon)p$, so $\mathcal{T} = \operatorname{logit} q$. At $p = 1$, $\mathcal{T} = \operatorname{logit}(1-\epsilon)$; at $p = 0.6$, $\epsilon = 0.1$, $\mathcal{T} = \operatorname{logit} 0.58 = 0.3228$.

rDPO. If a clean preference q is observed through a symmetric flip rate ϵ, so that $p = \epsilon + (1-2\epsilon)q$, then $\mathcal{T}_{\mathrm{rDPO}}(p) = \operatorname{logit} q$ recovers the clean logit and $\mathcal{T}_{\mathrm{cDPO}}$ applies the noise a second time. $\ell_\epsilon(\Delta) = [-(1-\epsilon)\log\sigma(\Delta) + \epsilon\log\sigma(-\Delta)]/(1-2\epsilon)$. Collecting terms, $(1-2\epsilon)R = -(p-\epsilon)\log\sigma(\Delta) - (1-p-\epsilon)\log\sigma(-\Delta)$. For $\epsilon < p < 1-\epsilon$ both coefficients are positive and $\mathcal{T} = \operatorname{logit}((p-\epsilon)/(1-2\epsilon))$; for $p \geq 1-\epsilon$ no finite margin minimises R. At $p = 1-\epsilon$ the second coefficient vanishes and R decreases towards 0 as $\Delta \to \infty$, and for $p > 1-\epsilon$ it is negative and $R \to -\infty$; symmetrically for $p \leq \epsilon$. At $p = 0.6$, $\epsilon = 0.1$, $\mathcal{T} = \log(0.5/0.3) = 0.5108$; at $p = 0.9$, $\mathcal{T} = +\infty$. Since a symmetric flip rate ϵ maps a clean q to $p = \epsilon + (1-2\epsilon)q$, $\mathcal{T}_{\mathrm{rDPO}}(p) = \operatorname{logit} q$, which is the sense in which rDPO is the unbiased correction (Chowdhury et al., 2024, Lemma 1) and cDPO is not.

Shifted logistic (SimPO). $\ell(\Delta) = -\log\sigma(\Delta - \gamma)$, so $\ell(-\Delta) = -\log\sigma(-\Delta - \gamma)$ and $R' = -p\,\sigma(\gamma - \Delta) + (1-p)\,\sigma(\Delta + \gamma)$. At $p = \frac{1}{2}$ the root is $\Delta = 0$; the map is increasing. Numerically at $\gamma = \frac{1}{2}$: $p = 0.6 \mapsto 0.533$, $0.7 \mapsto 1.092$, $0.8 \mapsto 1.732$, $0.9 \mapsto 2.628$.

Exponential. $R = pe^{-\Delta} + (1-p)e^{\Delta}$, $R' = 0$ at $e^{2\Delta} = p/(1-p)$, so $\mathcal{T} = \frac{1}{2}\operatorname{logit} p$.

Soft-label logistic with fixed z (DNO). With $p = 1$ on the recorded direction, $R' = \sigma(\Delta) - z$, minimiser $\operatorname{logit} z$; the target depends on the pair through z.

Numerical values at $\epsilon = 0.1$, $m = 1$, $\gamma = \frac{1}{2}$: for $p = (0.5, 0.6, 0.7, 0.8, 0.9, 0.99)$ the logistic map gives $(0, 0.405, 0.847, 1.386, 2.197, 4.595)$, squared $(0, 0.2, 0.4, 0.6, 0.8, 0.98)$, cDPO $(0, 0.323, 0.663, 1.046, 1.516, 2.111)$, rDPO $(0, 0.511, 1.099, 1.946, +\infty, +\infty)$, SimPO $(0, 0.533, 1.092, 1.732, 2.628, 5.089)$. Every value was obtained both from the closed form and by a grid minimisation of R over $[-30, 30]$; Figure 1b plots the first four.

Table 10: Table 5 with each score as a formula. "ref": F fixed, M mixture, T previous iterate, ∅ none. $\mathcal{E}$ marks extra terms. Greek letters are each method's own hyperparameters.

Method	Score $s(x, y)$	Link	ω scope	ref	$\mathcal{E}$	Rank	Loss
Margin objectives							
DPO	$\beta \log(\pi_\theta/\pi_{\text{ref}})$	logistic	const	F		A	A1
RSO (sigmoid)	$\gamma \log(\pi_\theta/\pi_{\text{sft}})$	logistic	const	F		A	A1
RSO (hinge)	$\gamma \log(\pi_\theta/\pi_{\text{sft}})$	hinge $m{=}1$	const	F		A	A2
cDPO	$\beta \log(\pi_\theta/\pi_{\text{ref}})$	label-smoothed	const	F		A	A3
rDPO	$\beta \log(\pi_\theta/\pi_{\text{ref}})$	noise-corrected	const	F		A	A4
IPO	$\log(\pi_\theta/\pi_{\text{ref}})$	squared	const	F		A	A5
IPO-MD	$\log(\pi_\theta/\pi_{\text{ref}})$	squared	const	F		A	A5
WPO	$\beta \log(\pi_\theta/\pi_{\text{ref}})$	logistic	pair (detached π_θ)	F		A	A6
XPO	$\beta \log(\pi_\theta/\pi_{\text{ref}})$	logistic	const	F	optimism	A	A7
f-DPO	$\beta f'(\pi_\theta/\pi_{\text{ref}})$	logistic	const	F		A	A8_f
INPO	$\log \pi - \frac{\tau}{\eta} \log \pi_{\text{ref}} - \frac{\eta-\tau}{\eta} \log \pi_t$	squared	const	M		B_t	B_t
DNO	$\eta \log(\pi/\pi_t)$	soft-label logistic	pair (z)	T		C_t	C_t
TDPO	$\beta \log(\pi_\theta/\pi_{\text{ref}}) + \beta D_{\text{SeqKL}}$	logistic	const	F		D	D
R-DPO	$\beta \log(\pi_\theta/\pi_{\text{ref}}) + \alpha\lvert y\rvert$	logistic	const	F		E	E
α-DPO	$\frac{\beta}{\lvert y\rvert} \log \pi_\theta - \text{sg}[\gamma + \alpha M^*]$	logistic	pair	M		F	F
SimPO	$\frac{\beta}{\lvert y\rvert} \log \pi_\theta$	logistic, shift γ	const	∅		G	G1
RRHF	$\frac{1}{\lvert y\rvert} \log \pi_\theta$	hinge $m{=}0$	const	∅	SFT	G	G2
ORPO	$\text{logit}\, P$, $P = e^{\log \pi_\theta/\lvert y\rvert}$	logistic	const	∅	SFT	G	G3
CPO	$\beta \log \pi_\theta$	logistic	const	∅	NLL	H	H1
SLiC-HF	$\log \pi_\theta$	hinge $m{=}\delta$	const	∅	CE on y_{ref}	H	H2
RankNet	$f(x_i)$	CE, soft $\bar{P}$	const	∅		I	I1
LambdaRank	$f(x_i)$	logistic	pair ($\lvert\Delta Z\rvert$)	∅		I	I2
Per-response objectives (no pairwise margin)							
KTO	$\sigma(\beta(\log \frac{\pi_\theta}{\pi_{\text{ref}}} - z_0))$	per-response	λ_D, λ_U; microbatch z_0	F		A	—
BCO	$\beta \log(\pi_\theta/\pi_{\text{ref}}) - \delta$	per-response BCE	EMA δ	F		A	—
APO (anchored)	$\beta \log(\pi_\theta/\pi_{\text{ref}})$	per-response σ		F		A	—
SPPO	$\log(\pi_\theta/\pi_t)$	per-response squared		T		J_t	—
Policy-gradient and token-level objectives (no pairwise margin, no per-response score)							
GRPO	advantage $(r - \mu_G)/\text{sd}_G$, clipped ratio, k_3 KL		batch	F			
RLOO	advantage r − LOO mean		batch (cancels)	F			
ReMax	advantage $r - r$(greedy)		prompt	F			
PPO-RLHF	advantage from a critic, clipped ratio			F	ptx		
Nash-MD-PG	$\nabla \log \pi_\theta \left[\mathcal{P}(y \succ y') - \frac{1}{2} - \tau \log \frac{\pi_\theta}{\mu}\right]$		pair	M			
RTO	token-level reward from a DPO policy, PPO			F			
Listwise (no pairwise decomposition for $n > 2$)							
ListNet	$f(x_j)$	softmax CE	item	∅		I	—
ListMLE	$g(x_j)$	Plackett–Luce NLL		∅		I	—

D PER-METHOD DERIVATIONS

Each entry gives the objective used (equation number and source file), the tuple $(s, \ell, \omega, \mathcal{E}, \mathcal{D})$, and the derivation of the row in Table 5. Objectives were taken from arXiv e-print sources; page references are given for the four non-arXiv papers. Numbers in square brackets are the arXiv identifiers. Acronym collisions are resolved by the objective, not the name: "APO" is Anchored Preference Optimisation (D'Oosterlinck et al., 2024); "f-DPO" is Wang et al. (2023); R-DPO (Park et al., 2024) and rDPO (Chowdhury et al., 2024) are different methods; RSO (Liu et al., 2024) defines two losses and is carded under both.

DPO (Rafailov et al., 2023) [2305.18290, eq. 7]. $-\mathbb{E} \log \sigma(\beta \log \frac{\pi_\theta(y_w)}{\pi_{\text{ref}}(y_w)} - \beta \log \frac{\pi_\theta(y_l)}{\pi_{\text{ref}}(y_l)})$. $s = \beta \log(\pi_\theta/\pi_{\text{ref}})$, logistic, $\omega = 1$, fixed reference, offline hard-label pairs. β multiplies the score and is a per-prompt-constant scale, so it does not affect Score.

RSO (Liu et al., 2024) [2309.06657, eqs. 8, 10]. Two losses on $\gamma[\log \frac{\pi_\theta}{\pi_{\text{sft}}}(y_w) - \log \frac{\pi_\theta}{\pi_{\text{sft}}}(y_l)]$: sigmoid-norm ($-\log \sigma$) and hinge-norm ($\max(0, 1 - \cdot)$). The sigmoid loss is DPO's loss with γ for β, hence Loss-equivalent to DPO; the method's contribution is the sampling scheme (statistical rejection sampling from a reward-induced policy), which changes $\mathcal{D}$ and not the loss. The hinge loss has a different link.

cDPO (Mitchell, 2023, eqs. 3–4). $(1 - \epsilon)\mathcal{L}_{\text{DPO}}(y_w, y_l) + \epsilon\mathcal{L}_{\text{DPO}}(y_l, y_w)$. DPO's score with the label-smoothed link; ϵ is global so ω is constant. Link not monotone (Appendix C).

rDPO (Chowdhury et al., 2024) [2403.00409, eq. 16]. $[(1 - \epsilon)\mathcal{L}(y_w, y_l) - \epsilon\mathcal{L}(y_l, y_w)]/(1 - 2\epsilon)$ with the DPO loss $\mathcal{L}$. Same score; link strictly decreasing for $\epsilon < \frac{1}{2}$; the minus sign and the $1/(1 - 2\epsilon)$ rescaling distinguish it from cDPO. TRL's `robust` branch implements this expression.

IPO (Azar et al., 2024) [2310.12036, eq. 17]. $\mathbb{E}[(h_\pi(y_w, y_l) - \frac{1}{2\tau})^2]$ with $h_\pi = \log \frac{\pi(y_w)\pi_{\text{ref}}(y_l)}{\pi(y_l)\pi_{\text{ref}}(y_w)}$. Score $\log(\pi/\pi_{\text{ref}})$, squared link with target $\frac{1}{2\tau}$. TRL's `ipo` branch length-normalises the scores before squaring, a modification the paper does not make.

IPO-MD (Calandriello et al., 2024) [2403.08635, eq. 14 and the Section 5.1 display]. The IPO loss with (Y, Y') drawn from $\text{sg}[\pi^{1-\beta}\pi_{\text{ref}}^{\beta}]$ (this β is the paper's mixture weight, not DPO's temperature). Same (s, ℓ, ω) as IPO; only $\mathcal{D}$ changes, with the reference inside the score fixed. The paper proves that the minimiser of online IPO is the Nash equilibrium of the regularised game and that IPO-MD(β) and Nash-MD-PG(β) share a fixed point, an Opt-level statement between a margin objective and a policy gradient (Figure 1a).

WPO (Zhou et al., 2024) [2406.11827, eqs. 1–2]. $-\mathbb{E}[w(y_w)w(y_l)\log p(y_w \succ y_l)]$ with w the detached, length-normalised policy probability. DPO's score and link; $\omega = w(y_w)w(y_l)$ is pair-dependent.

XPO (Xie et al., 2024) [2405.21046, Algorithm 1]. $\alpha \sum_i \log \pi(\tilde{\tau}^{(i)}) - \sum_{\mathcal{D}_{\text{pref}}} \log \sigma(\Delta_{\text{DPO}})$ with $\tilde{\tau}^{(i)} \sim \pi_{\text{ref}}$. The DPO term is certifiable; the optimism term is evaluated on single responses outside every pair and is unbounded below in the level of $\log \pi$, so it is in $\mathcal{E}$. At $\alpha = 0$ XPO is online DPO.

f-DPO (Wang et al., 2023) [2309.16240, eq. 3]. $-\mathbb{E}\log\sigma(\beta f'(\frac{\pi_\theta}{\pi_{\text{ref}}}(y_w)) - \beta f'(\frac{\pi_\theta}{\pi_{\text{ref}}}(y_l)))$. Score $\beta f'(\pi_\theta/\pi_{\text{ref}})$, logistic. For f the reverse KL, $f'(t) = 1 + \log t$ and the score is DPO's plus a constant; for JS, forward KL or α-divergences f' is a non-affine increasing function of the log-ratio, so the method is Rank-equivalent and not Score-equivalent to DPO. One label covers a family.

INPO (Zhang et al., 2024) [2407.00617, eq. 9]. $\mathbb{E}[(h_t(\pi, y_w, y_l) - \frac{1}{2\eta})^2]$ with h_t the difference of $\phi(y) = \log \pi(y) - \frac{\tau}{\eta}\log \pi_{\text{ref}}(y) - \frac{\eta-\tau}{\eta}\log \pi_t(y)$. A margin objective with the squared link and a mixture reference; no Bradley–Terry model is assumed. Score-equivalent to IPO only at $t = 1$ when $\pi_1 = \pi_{\text{ref}}$.

DNO (Rosset et al., 2024) [2404.03715, eq. 7]. $\max_\pi \mathbb{E}\{z \log\sigma(\Delta) + (1-z)\log\sigma(-\Delta)\}$ with $\Delta = \eta[\log \frac{\pi}{\pi_t}(y_1) - \log \frac{\pi}{\pi_t}(y_2)]$ and $z = \sigma(r_t(y_1) - r_t(y_2))$. Score against the previous iterate; soft-label logistic link whose stationary margin is $\text{logit}\, z$ per pair, so ω is pair-dependent through z and Loss fails by the pair-weight witness. DPO is $z \equiv 1$ with $\pi_t = \pi_{\text{ref}}$.

TDPO (Zeng et al., 2024) [2404.11999, eqs. 25, 27]. $-\mathbb{E}\log\sigma(u - \delta)$ with u the DPO margin and δ a difference of per-response sequential forward KLs $\beta D_{\text{SeqKL}}(x, y; \pi_{\text{ref}} \| \pi_\theta)$. The effective score is $\beta\log(\pi_\theta/\pi_{\text{ref}}) + \beta D_{\text{SeqKL}}$, a per-response quantity, so a single scalar margin remains; TDPO$_2$ applies a stop-gradient to the chosen-response term.

R-DPO (Park et al., 2024) [2403.19159, eq. 9]. DPO with $\alpha|y_w| - \alpha|y_l|$ added inside the logit. Score $\beta\log(\pi_\theta/\pi_{\text{ref}}) + \alpha|y|$; the cleanest instance of an additive per-response shift.

α-DPO (Wu et al., 2024a) [2410.10148, eq. 12]. $-\mathbb{E}\log\sigma(u - \text{sg}[\gamma + \alpha M^*])$ with u SimPO's margin, M DPO's margin and M^* its Z-score over the training set. The stop-gradient makes the effective margin depend on dataset statistics (μ_M, σ_M), a population-level analogue of GRPO's normalisation; recorded as pair-dependent ω.

SimPO (Meng et al., 2024) [2405.14734, eq. 11]. $-\mathbb{E}\log\sigma(\frac{\beta}{|y_w|}\log\pi_\theta(y_w) - \frac{\beta}{|y_l|}\log\pi_\theta(y_l) - \gamma)$. Score $(\beta/|y|)\log\pi_\theta$, logistic with shift γ folded into the link, no reference.

RRHF (Yuan et al., 2023) [2304.05302, eqs. 2–5]. $\sum_{r_i<r_j}\max(0, p_i - p_j) + L_{\text{ft}}$ with p_i the length-normalised log-likelihood and L_{ft} the NLL of the best-rewarded response. Score $|y|^{-1}\log\pi_\theta$ (SimPO's up to β), zero-margin hinge, SFT term in $\mathcal{E}$.

ORPO (Hong et al., 2024) [2403.07691, eqs. 3–7]. $\mathbb{E}[L_{\text{SFT}} + \lambda L_{\text{OR}}]$, $L_{\text{OR}} = -\log\sigma(\log\frac{\text{odds}(y_w)}{\text{odds}(y_l)})$, $\text{odds} = P/(1-P)$, $\log P = |y|^{-1}\log\pi_\theta$. Score $\text{logit}\, P$, logistic, SFT term in $\mathcal{E}$. Rank-equivalent to SimPO and RRHF; the non-affine witness (per-token log-probabilities $(-0.1, -0.2)$ against $(-3.0, -3.1)$, both with $\Delta_{\text{SimPO}} = 0.1$ at $\beta = 1$, giving $\Delta_{\text{ORPO}} = 0.744$ and 0.105) separates it at Score.

CPO (Xu et al., 2024a) [2401.08417, eq. 5]. $\mathcal{L}(\pi_\theta; U) - \mathbb{E} \log \pi_\theta(y_w)$ with $\mathcal{L}(\pi_\theta; U)$ the DPO loss under a uniform reference. Score $\beta \log \pi_\theta$ (reference-free, not length-normalised, so Score-equivalent to SLiC-HF's score and not to SimPO's), logistic, NLL term in $\mathcal{E}$.

SLiC-HF (Zhao et al., 2023) [2305.10425, eq. 4]. $\max(0, \delta - \log P_\theta(y^+) + \log P_\theta(y^-)) - \lambda \log P_\theta(y_{\text{ref}})$. Score $\log \pi_\theta$, hinge with margin δ, regulariser on a third sequence in $\mathcal{E}$.

RankNet (Burges et al., 2005, eqs. 1–3, p. 3). $C_{ij} = -\bar{P}_{ij} o_{ij} + \log(1 + e^{o_{ij}})$ with $o_{ij} = f(x_i) - f(x_j)$. Generic scorer; with hard $\bar{P}$ the logistic link, with the soft targets the paper supports the cDPO link.

LambdaRank (Burges et al., 2006, eq. 3), (Burges, 2010, Sec. 7). $C = \sum |\Delta Z_{ij}| \log(1 + e^{-\sigma(s_i - s_j)})$. Generic scorer, logistic link, $\omega = |\Delta Z_{ij}|$ pair-dependent and list-dependent.

ListNet (Cao et al., 2007, eq. 2, p. 5), **ListMLE** (Xia et al., 2008, eq. 9, p. 4). A cross-entropy between top-one softmax distributions and a Plackett–Luce negative log-likelihood respectively. Both have a per-item score and neither decomposes into pairwise margins for $n > 2$; ListNet reduces to RankNet at $n = 2$ by the paper's own statement.

KTO (Ethayarajh et al., 2024) [2402.01306, eq. 8]. $\mathbb{E}[\lambda_y - v(x, y)]$ with $v = \lambda_D \sigma(\beta(r_\theta - z_0))$ on desirable and $\lambda_U \sigma(\beta(z_0 - r_\theta))$ on undesirable examples, $r_\theta = \log(\pi_\theta / \pi_{\text{ref}})$, z_0 a KL estimated on the microbatch (clamped at zero). Per-response; a paired contribution $\lambda_D(1 - \sigma(\beta(r_w - z_0))) + \lambda_U(1 - \sigma(\beta(z_0 - r_l)))$ is a function of r_w and r_l separately.

BCO (Jung et al., 2024) [2404.04656, eqs. 9–10]. $\mathbb{E}_{\mathcal{D}^+}[-\log \sigma(r_\theta - \delta)] + \mathbb{E}_{\mathcal{D}^-}[-\log \sigma(-(r_\theta - \delta))]$ with δ the average of the mean rewards of the two sets, tracked by an EMA. Per-response, with a dataset-level shift.

APO (Anchored) (D'Oosterlinck et al., 2024) [2408.06266, eqs. 6–7]. $\mathcal{L}_{\text{zero}} = -\sigma(r_w) + \sigma(r_l)$, $\mathcal{L}_{\text{down}} = \sigma(r_w) - \sigma(r_w - r_l)$. Neither is a function of $r_w - r_l$ alone. Two other papers share the acronym, Active PO [2402.10500], a Bradley–Terry logistic loss on linear features with an active query rule, and Adversarial PO [2311.08045], a min-max game; the classification of the row depends on the resolution, which is why the name is fixed here.

SPPO (Wu et al., 2024b) [2405.00675, eqs. 11, 13]. Proposition 6. TRL's `sppo_hard` branch is a sum of two per-response squared terms against π_{ref}; the paper uses π_t.

GRPO (Shao et al., 2024) [2402.03300, eqs. 3–4]. Clipped-ratio surrogate with $\hat{A}_{i,t} = (r_i - \text{mean}(\mathbf{r}))/\text{std}(\mathbf{r})$ and the k_3 KL estimator added to the loss. Whether std is the sample or population deviation is not stated; TRL v1.13.0 uses the sample deviation.

RLOO (Ahmadian et al., 2024) [2402.14740, eq. RLOO]. $\frac{1}{k} \sum_i [R(y_i) - \frac{1}{k-1} \sum_{j \neq i} R(y_j)] \nabla \log \pi(y_i)$; the KL penalty is inside R. No importance ratio, no clipping, no standard deviation.

ReMax (Li et al., 2024) [2310.10505, eqs. 11–12]. REINFORCE with the baseline $r(x, \bar{a}_{1:T})$ of the greedy rollout; prompt-level, not batch-level.

PPO-RLHF (Ouyang et al., 2022) [2203.02155, eq. 2], (Schulman et al., 2017) [1707.06347, eq. 7]. Clipped surrogate on a learned advantage with a per-token KL penalty inside the reward and the pretraining-mixture term.

Nash-MD-PG (Munos et al., 2024) [2312.00886, eqs. 16, 18]. $\nabla \log \pi_\theta(y)[\mathcal{P}(y \succ y') - \frac{1}{2} - \tau \log \frac{\pi_\theta(y)}{\mu(y)}]$ with y' from the geometric mixture $\pi_\theta^{1-\beta} \mu^\beta$ (β the paper's mixture weight). The tabular Nash-MD (eq. 6) is an argmax over policies; the implemented method is the gradient estimate.

RTO (Zhong et al., 2025) [2404.18922, eq. 4.9]. Token-level rewards $\beta_1 \log \frac{\pi_{\text{dpo}}}{\pi_{\text{ref}}}(y_h \mid \cdot) - \beta_2 \log \frac{\pi}{\pi_{\text{ref}}}(y_h \mid \cdot)$ with a terminal sentence-level term, optimised by PPO.

E Experimental details and all checkpoints

Scores. With the implicit reward $r = \log \pi_\theta - \log \pi_{\text{ref}}$: DPO βr; IPO r; SLiC $\log \pi_\theta$; SimPO $(\beta/|y|) \log \pi_\theta$ with $\beta = 2$; ORPO $\text{logit} \exp(\log \pi_\theta / |y|)$; KTO $\sigma(\beta(r - z_0))$ with z_0 the mean of r over the scored candidates, clamped at zero. The SimPO/ORPO residual in Table 6 is small

because at this model's operating point the per-token likelihood P is far from one, so $\mathrm{logit}\, P = \log P - \log(1-P)$ is nearly linear in $\log P$; the non-affine witness of Section 3 lives at $P \approx 0.9$, which these candidates do not reach.

Configuration. Device Apple M5 (10-core GPU, 32 GB), PyTorch 2.14 MPS backend, float32; transformers 5.17, datasets 5.0. Probe set: the first 500 prompts of the `train_prefs` split, four responses per prompt sampled from π_{ref} at temperature 0.8 with 256 new tokens and seed 0, plus the dataset's chosen and rejected responses. Prompts are cut to 512 tokens before sampling, so for the 37 prompts longer than this the sampled responses continue the prompt text. Training pairs are `train_prefs` rows 500 onward. Optimiser AdamW; checkpoints at steps 0, 100, 200 and 300, with step 0 excluded from the tables because $r \equiv 0$ there. Scoring truncates the prompt and the response at 256 tokens each. This cuts 103 of the 500 probe prompts and 410 of the 1 000 dataset responses, and every method's score is computed from the same truncated sequence, so all methods are compared on identical inputs. The KTO reference point z_0 estimated to zero at every checkpoint, so its score reduced to $\sigma(\beta r)$. The pair-level weights λ_D, λ_U are not part of the score and are not measured.

Main configuration (Tables 6 and 11). Effective batch 32, learning rate 5×10^{-6}, 300 steps, training sequences of 256 prompt and 256 completion tokens, three seeds.

Near-reference configuration (Table 12). Effective batch 8 (micro-batch 4, two accumulation steps), learning rate 5×10^{-7}, 300 steps, training sequences truncated to 128+128 tokens, seed 0, 2 400 training pairs consumed. At this learning rate the policy stays close to π_{ref}, and cross-boundary agreement is within noise of the random control at every checkpoint. The certified identities are exact here as well.

Table 11: Main configuration (Table 6): sign agreement over the 3 000-candidate probe set by checkpoint, mean ± sd over three seeds. Mean $|r|$ is the mean absolute implicit reward over the candidates.

Pair	step 100	step 200	step 300
mean $\lvert r \rvert$	15.55±1.00	19.30±0.64	22.10±1.45
DPO vs rearranged (control)	1.000±0.000	1.000±0.000	1.000±0.000
DPO vs random (control)	0.500±0.018	0.504±0.004	0.501±0.003
DPO vs IPO	1.000±0.000	1.000±0.000	1.000±0.000
DPO vs KTO score	1.000±0.000	1.000±0.000	1.000±0.000
SimPO vs ORPO	1.000±0.000	1.000±0.000	1.000±0.000
SLiC vs SimPO	0.753±0.009	0.744±0.005	0.740±0.005
DPO vs SimPO	0.573±0.008	0.592±0.009	0.593±0.003
DPO vs ORPO	0.573±0.008	0.592±0.009	0.593±0.003
DPO vs SLiC	0.596±0.016	0.631±0.012	0.630±0.011

Table 12: Near-reference configuration (learning rate 5×10^{-7}, batch 8, one seed), all checkpoints. Entries are ρ / sign agreement / τ_{K}.

Pair	step 100	step 200	step 300
DPO vs rearranged	0.000/1.000/1.000	0.000/1.000/1.000	0.000/1.000/1.000
DPO vs random	0.906/0.494/−0.012	0.907/0.500/0.000	0.887/0.510/0.020
DPO vs IPO	0.000/1.000/1.000	0.000/1.000/1.000	0.000/1.000/1.000
DPO vs KTO	0.008/1.000/1.000	0.006/1.000/1.000	0.014/1.000/1.000
SimPO vs ORPO	0.006/1.000/1.000	0.005/1.000/1.000	0.005/1.000/1.000
SLiC vs SimPO	0.482/0.784/0.571	0.489/0.780/0.563	0.500/0.773/0.547
SLiC vs ORPO	0.483/0.784/0.571	0.489/0.780/0.563	0.500/0.773/0.547
DPO vs SimPO	0.859/0.516/0.035	0.814/0.532/0.066	0.805/0.538/0.080
DPO vs ORPO	0.861/0.516/0.035	0.821/0.532/0.066	0.812/0.538/0.080
DPO vs SLiC	0.785/0.546/0.096	0.816/0.504/0.012	0.817/0.505/0.014
IPO vs KTO	0.008/1.000/1.000	0.006/1.000/1.000	0.014/1.000/1.000
SLiC vs KTO	0.783/0.546/0.096	0.815/0.504/0.012	0.817/0.505/0.014
SimPO vs KTO	0.860/0.516/0.035	0.817/0.532/0.066	0.808/0.538/0.080

E.1 Experiment 3: details

Configuration. Same initialisation and data order as Experiment 1, three seeds. The soft label is $p = \sigma(\kappa(s_c - s_r))$ with s_c, s_r the per-response ratings that `ultrafeedback_binarized` carries for the chosen and rejected response, and κ fixed so that the mean of p on the training set is 0.75. IPO uses $\tau = \frac{1}{2}$, so the squared link's target is $m = 1$. Batch 8, learning rate 5×10^{-6}, 200 steps. DPO's mean probe $|\Delta|$ is 0.995 at step 100 and 1.186 at step 200. The first bin of Table 13 holds only tied pairs ($p = \frac{1}{2}$, every target zero), so the comparisons in Section 6 use the three bins in $(0.5, 0.9)$.

Additional statistics. The 9 adjacent differences have z from 1.8 to 6.5. For DPO/cDPO, DPO/rDPO and cDPO/rDPO the through-origin fit alone has $R^2 \geq 0.98$, and the fraction of target reached (26% to 31%) is the through-origin slope. IPO's largest bin mean is 0.34 and its bin means are not linear in p ($R^2 = 0.83$). The ratio rDPO/DPO is 1.71, 1.48 and 1.28 across the three bins, falling with p where the target maps have it rising (1.26 at $p = 0.6$, 1.40 at $p = 0.8$).

Table 13: Experiment 3: held-out bin means of Δ at step 200 (mean over 3 seeds, $\pm$ standard error over pairs of the seed-averaged Δ), with each method's target $\mathcal{T}_\ell$ at the bin's mean p in parentheses.

Method	$p \in [0.5, 0.6)$	$p \in [0.6, 0.7)$	$p \in [0.7, 0.8)$	$p \in [0.8, 0.9)$	$p \in [0.9, 1.0]$
mean p (n)	0.500 (272)	0.646 (764)	0.772 (229)	0.849 (268)	0.966 (467)
DPO	-0.08 ± 0.06 (0.00)	$+0.11 \pm 0.03$ (0.60)	$+0.35 \pm 0.07$ (1.22)	$+0.54 \pm 0.07$ (1.73)	$+0.81 \pm 0.06$ (3.34)
IPO	$+0.00 \pm 0.04$ (0.00)	$+0.03 \pm 0.03$ (0.29)	$+0.06 \pm 0.05$ (0.54)	$+0.19 \pm 0.05$ (0.70)	$+0.34 \pm 0.04$ (0.93)
cDPO	-0.05 ± 0.05 (0.00)	$+0.07 \pm 0.03$ (0.48)	$+0.23 \pm 0.06$ (0.93)	$+0.41 \pm 0.06$ (1.26)	$+0.62 \pm 0.05$ (1.93)
rDPO	-0.12 ± 0.08 (0.00)	$+0.18 \pm 0.04$ (0.77)	$+0.52 \pm 0.09$ (1.66)	$+0.70 \pm 0.09$ (2.70)	$+1.19 \pm 0.09$ (∞)

E.2 Experiment 2 protocol and sweep

Instances of Table 7. Chain margins are sums of $g = \mathcal{T}_\ell \circ F$ over the steps; star margins anchored at the worst response y_w are $g(u_1 - u_w) - g(u_K - u_w)$. Bradley–Terry, utilities $(5, 0, 1, 2, 3)$, steps $(+5, -1, -1, -1)$, anchor y_2: DPO $5 - 3 = 2$; IPO $0.987 - 3 \times 0.462 = -0.400$, star $0.987 - 0.905 = 0.081$; cDPO $\operatorname{logit}(0.1 + 0.8\,\sigma(5)) - 3\,\operatorname{logit}(0.1 + 0.8\,\sigma(-1)) = -0.189$, star 0.307. Thurstone, utilities $(0, 4, 3, 2, 1, -1)$, steps $(-4, +1, +1, +1, +2)$, anchor y_6 so the star margin is $g(1)$: DPO $\operatorname{logit}\Phi(-4) + 3\,\operatorname{logit}\Phi(1) + \operatorname{logit}\Phi(2) = -1.595$, star $\operatorname{logit}\Phi(1) = 1.668$; IPO 2.003, star 0.683; cDPO 3.490, star 1.226. Every entry is recomputed by `exp2/check_table.py` in the supplementary material.

Sweep. For $K \in \{4, 6, 8\}$ and N draws of $u_i \sim \mathcal{N}(0, \sigma_u^2)$, $\sigma_u \in \{1, 2\}$: generators Bradley–Terry, Thurstone and Bradley–Terry with flip rate 0.1; designs rank-adjacent chain, random-order chain, star at the worst response, star at a random response, random spanning tree, complete; links DPO, IPO ($m = 1$), cDPO ($\epsilon = 0.1$), rDPO ($\epsilon = 0.1$, only where every labelled $p_{ij} \in (\epsilon, 1 - \epsilon)$). Tree designs are solved in closed form by Proposition 4; the interpolation that gives unlabelled pairs weight $w \in \{0, 0.01, 0.03, 0.1, 0.3, 1\}$ is a convex problem solved numerically. Reported per cell: reversal rate on unlabelled pairs, rate at which the best response stays best, Kendall τ_{K} between fitted scores and u.

Solver and checks. Interpolation and complete-design problems are solved per draw by damped Newton with the exact Hessian until every draw's largest gradient component is below 10^{-12}; on 100 random problems scipy's L-BFGS-B (tolerance 10^{-10}) started from zero and from a random point agrees with Newton to 4×10^{-8}. `exp2/run_exp2.py` asserts, in every cell, that stars, complete designs and chains in utility order reverse no pair and that chains in any order reverse no pair when $\mathcal{T} \circ F$ is linear (2 394 assertions, none failed). A pair counts as reversed only if its fitted margin has the wrong sign; exact ties from the floating-point saturation of Φ beyond $|d| \approx 8$ (at most 181 of the 210 000 evaluated pairs of a cell) count as neither. rDPO is defined on a draw only when every pair that carries weight has $p_{ij} \in (0.1, 0.9)$, and its entries are over those draws.

Results. Table 14 gives the slice $K = 6$, $\sigma_u = 2$ at $w = 0$ and Figure 4 the interpolation; every cell is in `exp2/results/sweep.csv`. In every one of the six (K, σ_u) cells, on chains in random order and random spanning trees, the only links that reverse no pair are DPO under Bradley–Terry labels and rDPO under Bradley–Terry labels with flip rate 0.1, the two with $\mathcal{T} = F^{-1}$ (rDPO's de-noising inverts the flip); every other link–generator pair reverses some pairs, as Proposition 4(b) requires.

Under flipped labels DPO's map $\mathrm{logit}(0.1 + 0.8\,\sigma(d))$ equals cDPO's under clean labels, and the two reverse the same pairs.

Table 14: Experiment 2 sweep ($K = 6$, $\sigma_u = 2$, 10 000 draws, $w = 0$): percentage of unlabelled pairs reversed (complete design: all pairs). Bootstrap standard errors ≤ 0.12 points. rDPO columns use only draws with every labelled $p_{ij} \in (0.1, 0.9)$.

	Bradley–Terry				Thurstone				BT, flip rate 0.1			
Design	DPO	IPO	cDPO	rDPO	DPO	IPO	cDPO	rDPO	DPO	IPO	cDPO	rDPO
chain, utility order	0	0	0	0	0	0	0	0	0	0	0	0
chain, random order	0	9.2	7.1	3.2	4.8	12.0	10.7	3.0	7.1	9.2	8.2	0
random spanning tree	0	7.3	5.5	2.5	3.6	9.6	8.5	1.9	5.5	7.3	6.4	0
star, worst response	0	0	0	0	0	0	0	0	0	0	0	0
star, random response	0	0	0	0	0	0	0	0	0	0	0	0
complete	0	0	0	0	0	0	0	0	0	0	0	0

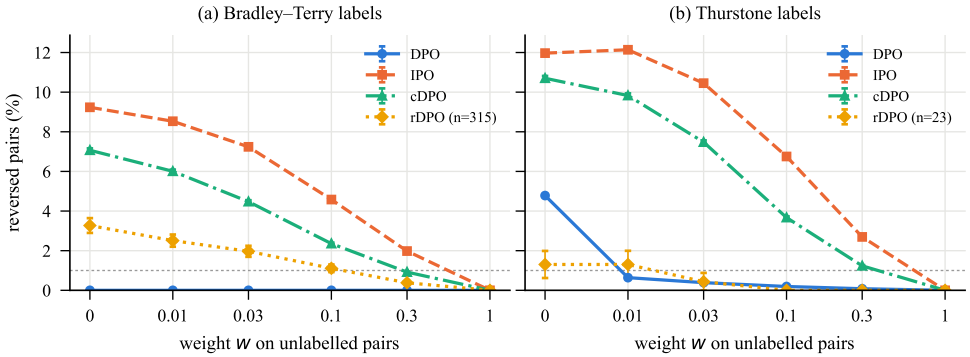

Figure 4: Reversed unlabelled pairs on random-order chains against the weight w on unlabelled pairs ($w = 0$ chain, $w = 1$ complete design). rDPO uses only draws where it is defined. Bars are bootstrap standard errors; the dotted line is 1%.

F Further related work

Pair construction. Chen et al. (2026) show for IPO that instance-dependent sampling strengthens ranking guarantees, and Experiment 2 is the complementary case of curated sparse designs. Liu et al. (2026) aggregate over preference graphs built from rollout rankings, and Proposition 4 says when the graph's shape decides the ranking. Shah & Wainwright (2018) and Rajkumar & Agarwal (2014) give the rank-aggregation theory to which Proposition 4(c) reduces.

Implicit bias and overoptimisation. Under hard labels a link with an unbounded target map has no finite minimiser, and the implicit-bias results for monotone losses (Soudry et al., 2018; Lyu & Li, 2020; Ji & Telgarsky, 2020) describe where gradient descent goes in that case. Overoptimisation curves (Gao et al., 2023; Rafailov et al., 2024a) compare distinct objectives at matched KL and are where Loss-distinct members of one Score class should next be compared.

Learning dynamics. Razin et al. (2024) show that DPO can lower the likelihood of both responses in a pair, Im & Li (2024) analyse how the distinguishability of preferences sets the speed of alignment, Tajwar et al. (2024) find that on-policy sampling and negative gradients matter across objectives, and Korbak et al. (2022) read KL-regularised RL as Bayesian inference. These results concern trajectories under one objective, and the levels here concern relationships between objectives. Fisch et al. (2025) supply a pair that shares an unconstrained optimum with different loss surfaces, used in Proposition 1(d). Chen et al. (2024) measure the ranking accuracy that Rank is about and find it low for trained models.

GRPO and policy gradients. Vojnovic & Yun (2025) characterise the stationary policy of the GRPO update, Liu et al. (2025b) remove its standard-deviation divisor as a bias correction, Mroueh (2025) rewrite it with binary rewards as a contrastive loss, and Wu et al. (2025) read the group statistic as a control variate. RLOO (Ahmadian et al., 2024) and ReMax (Li et al., 2024) use baselines that leave a

constant scale or none. Nash-MD (Munos et al., 2024), IPO-MD (Calandriello et al., 2024), INPO (Zhang et al., 2024), DNO (Rosset et al., 2024) and SPPO (Wu et al., 2024b) target a game-theoretic solution, and Appendix D records which of them are margin objectives.

Learning to rank. RankNet (Burges et al., 2005), LambdaRank (Burges et al., 2006; Burges, 2010), ListNet (Cao et al., 2007) and ListMLE (Xia et al., 2008) are included in Table 5 because preference optimisation inherits their pairwise and listwise forms. RankNet with soft targets has the cDPO link, and LambdaRank's list-dependent pair weight is the mechanism by which WPO and DNO fail Loss.