AGREEMENT IS NOT COVERAGE: WHAT EVALUATION CRITERIA REGISTER OF A TRAINING CHANGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Criteria that monitor optimisation are usually chosen because they agree with human judgements, or with one another, on existing responses. Such agreement does not show whether a criterion registers the changes that training makes. We call this difference the *agreement–coverage gap* and study it with linear criterion readouts on a shared representation, where the part of a training direction that a readout can register is its overlap with that readout and can be computed before training. Training Qwen2.5-1.5B with DPO along an unmodified reward-model direction already moves eleven FLASK criteria unevenly, in the order of their overlap (Spearman $\rho = 0.71$). In a controlled test, we remove from a driver criterion's direction only the component shared with a checker criterion and compare two DPO arms at matched KL. Across five pairs whose readouts rank 65%–80% of response pairs alike, the checker's detection of the policy change falls in every pair (aggregate $+0.063$, 95% CI $[0.060, 0.067]$) while the driver keeps 79%–99% of its above-chance signal. An equal-length placebo removal has a smaller effect in all three pairs where it could be matched, and pre-training overlap orders the responses of the other ten criteria ($\rho = 0.81$). The pattern holds at 0.5B and 7B and with a mis-specified checker readout. On generated text, an independent judge sees both criteria move, so our claims concern criterion readouts.

1 INTRODUCTION

Optimising a language model usually involves more than one evaluation criterion. One signal drives training, and other criteria are kept as checks on what the optimisation does (Bai et al., 2022; Ye et al., 2024; Wang et al., 2024b). A check is typically trusted because it agrees with human judgements, gold references or other criteria on existing responses (Zheng et al., 2023; Liu et al., 2023). Such agreement describes how criteria rank a fixed set of outputs, whereas a check is kept to notice how the model changes, and there is little reason to expect the two properties to coincide.

Evidence that they come apart has accumulated from several directions. Proxy and gold rewards diverge under optimisation (Gao et al., 2023), reward models with similar in-distribution accuracy disagree once a policy is aligned to them (Eisenstein et al., 2024), and evaluator LLMs miss many deliberate drops in quality (Doddapaneni et al., 2024). When optimisation pressure acts against a monitor, behaviour can move where the monitor does not look. Attacks can induce activations that evade latent-space probes while keeping the monitored behaviour (Bailey et al., 2025), and reasoning models trained against a chain-of-thought monitor learn to hide misbehaviour from it (Baker et al., 2025). These studies observe the failure after training. We ask whether it can be anticipated before training from the criteria themselves, and call the difference between ranking responses alike and registering the same training change the *agreement–coverage gap*. It is the evaluation-side counterpart of the drift paradox in fine-tuning, where updates of equal magnitude cause different damage because they point in different directions (Liu et al., 2026).

For criteria read out linearly from a shared representation, the gap can be computed. A policy update moves a readout's ranking of response pairs through the update's overlap with that readout, so every criterion's share of a training direction is available on the base model (Section 3). Figures 1 and 2 turn this into a controlled comparison. We take a driver criterion A and a checker B and remove from A's direction only the part that B's readout sees. The resulting direction $u_{A \setminus B}$ is orthogonal to B's readout coefficients and stays as close to A as that constraint allows. Orthogonality

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.

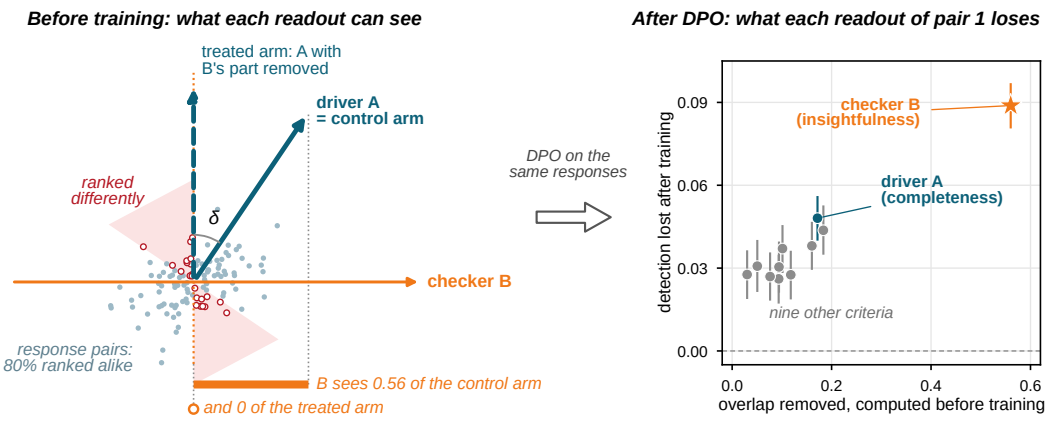

Figure 1: **Two readouts that rank responses alike can register different amounts of the same training change.** Left: geometry of pair 1 (`completeness→insightfulness`) drawn to scale. Points are response pairs (schematic); the two readouts rank 80% of them alike. The control arm trains along driver A; the treated arm removes the part of A that checker B's readout sees, so B's readout sees none of it while 83% of the alignment with A remains. Right: measured after DPO at matched KL, for all eleven readouts of the same pair. The horizontal axis is the overlap each readout loses by the construction, computed before training (0.56 for B, as on the left); the vertical axis is the detection it loses, with 95% question-bootstrap intervals.

in coefficient space is guaranteed by the construction; whether it survives DPO training, and at what cost to A, is what the experiments measure.

Coverage turns out to be uneven even before anything is removed. When Qwen2.5-1.5B is trained with DPO along a reward-model direction, the eleven FLASK criteria register the change roughly in the order of their overlap with that direction (Section 4.2). In the controlled comparison, five criterion pairs whose readouts rank 65%–80% of response pairs alike all show lower checker detection in the treated arm, while the driver keeps most of its signal (Table 2). The overlap computed before training also orders how the ten remaining criteria respond (Figure 4), and the checker's loss persists under a placebo removal of equal size, at two further model scales, with a mis-specified checker readout and under two external reward signals (Figure 6).

All of this concerns fixed linear readouts trained with DPO. When an LLM judge scores generated text, both criteria of a pair move together, and the overlap orders criteria within a comparison without predicting effect sizes across pairs (Section 5). For choosing checkers the finding is still useful, because response agreement is silent about which training changes a checker will register, while overlap can be computed in advance. Our contributions are:

- **The agreement–coverage gap, made computable.** A closed-form, checker-orthogonal training direction whose overlap with every criterion is available on the base model (Section 3).
- **Controlled evidence at matched displacement.** Selective detection loss in five of five pairs with the driver largely preserved, together with a placebo control, scale and readout variations and external reward signals (Sections 4.3 and 4.5).
- **Overlap orders what criteria register.** The same pre-training overlap orders criterion responses under an unmodified reward direction, the constructed directions and an external signal, and we report where this prediction stops (Sections 4.4 and 5).

2 Related Work

Separately trained preference models are routinely used both to drive and to evaluate RLHF, and their robustness failures are correlated (Bai et al., 2022). Proxy–gold divergence under optimisation has been quantified (Gao et al., 2023), ensembles have been proposed as conservative objectives

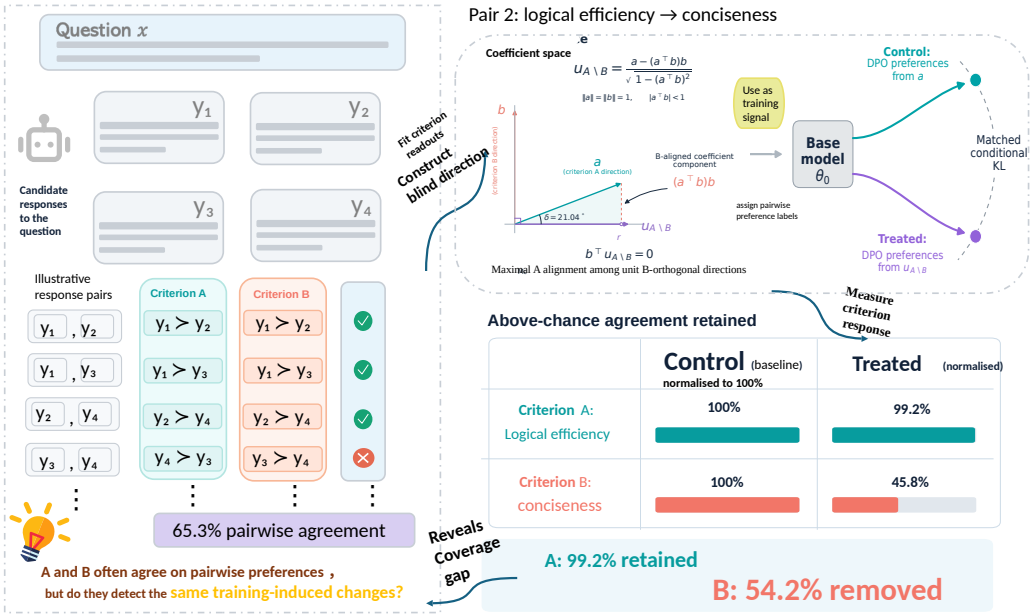

Figure 2: **Pipeline, shown for pair 2 (`logical efficiency`→`conciseness`).** Left: the readouts of the two criteria order 65.3% of same-question response pairs alike. Middle: in the coefficient space of the readouts, the blind direction $u_{A\setminus B}$ removes from a its component along b and is the unit direction orthogonal to b closest to a ($\delta = 21.04°$). Right: the control arm is trained with DPO on preferences labelled by a, the treated arm on preferences labelled by $u_{A\setminus B}$, and the two are compared at matched conditional KL. Bottom: detection above chance in the treated arm as a fraction of the control arm. The driver keeps 99.2% (R_A) and the checker loses 54.2% (E_B; Table 2).

(Coste et al., 2024), and reward models with similar in-distribution accuracy diverge after alignment while sharing errors (Eisenstein et al., 2024). Monitors evaded under optimisation pressure include latent-space probes (Bailey et al., 2025) and chain-of-thought monitors (Baker et al., 2025), and evaluator LLMs miss targeted quality drops (Doddapaneni et al., 2024). These works measure divergence or evasion once training has happened. We compute a checker's coverage of a training direction beforehand and test its selective loss at matched KL. Liu et al. (2026) showed that equal update magnitudes can cause different damage in fine-tuning; the agreement–coverage gap is the same distinction seen from the evaluation side.

LLM judges are validated by their agreement with human preferences (Zheng et al., 2023; Liu et al., 2023), and FLASK, HelpSteer2 and ArmoRM provide fine-grained attribute signals (Ye et al., 2024; Wang et al., 2024b;a). Such validation concerns a fixed set of responses. On the representation side, concept directions are extracted and steered (Zou et al., 2023), nullspace projection removes linearly decodable information (Ravfogel et al., 2020), and amnesic probing removes a property from the representation to test whether behaviour uses it (Elazar et al., 2021). Our intervention leaves the representation unchanged. It removes a component from one criterion's readout direction and asks whether a second criterion still registers the training that follows.

3 Coverage and a checker-orthogonal direction

We fit ridge probes of criterion scores on frozen representations of Qwen2.5-1.5B, taking h as the mean hidden state over response tokens at layer 17 of 28. With μ the fitting-set mean and D the diagonal matrix of fitting-set scales, the readout for criterion X is $s_X(h) = [D^{-1}(h-\mu)]^\top W_X$. Two readouts show *response agreement* on a same-question response pair when their score differences have the same sign. This is the property usually checked when a criterion is validated, and for the five pairs studied here it holds on 65%–80% of response pairs (Table 1).

Table 1: The five pre-registered criterion pairs, frozen before training. acc_A and acc_B are the probes' held-out pairwise accuracies on SELECT, used for eligibility and for the orientation rule. C2 is the fraction of same-question response pairs on which the two readouts' score differences agree in sign, a selection condition requiring at least 0.60. c is the removable component and $\delta = \arcsin c$ the angle between the control direction u_A and the treated direction $u_{A\setminus B}$. Pair selection used acc and C2; it did not use c or any quantity predicting the outcome.

#	driver A	checker B	acc_A	acc_B	C2	c	δ
1	completeness	insightfulness	0.723	0.709	0.801	0.5602	34.07°
2	logical efficiency	conciseness	0.689	0.682	0.653	0.3590	21.04°
3	logical robustness	comprehension	0.676	0.649	0.741	0.5837	35.71°
4	logical correctness	commonsense understanding	0.629	0.628	0.761	0.6209	38.38°
5	readability	factuality	0.625	0.615	0.687	0.4352	25.80°

A check, in contrast, has to register a training change. Training moves the policy while the readouts stay fixed, so for each same-question response pair we compare the change in mean response-token log-probability since the base model with the sign of the readout's score difference. The fraction of pairs on which the two agree is the *detection* α_X of the training change by criterion X; the appendices call it the agreement statistic. Pairing within a question makes the question the unit of independence. About 6.5% of the response pairs consist of duplicate responses that every readout scores as tied, which places the chance level at $q_X \simeq 0.468$ rather than 0.5 (Appendix G).

To first order, a policy displacement d changes the log-probability margin of response pair i by $g_i^\top d$, where g_i is the margin gradient. With $q_{X,i} = s_X(h_{i1}) - s_X(h_{i2})$ the fixed readout margin of the same pair,

$$\Delta m_i = g_i^\top d + O(\|d\|^2), \qquad \text{criterion } X \text{ registers pair } i \iff q_{X,i}\, g_i^\top d > 0 \tag{1}$$

whenever the linear term dominates the remainder (Appendix Y.2). A training direction that labels preferences by a readout thus reaches each criterion through its overlap with that criterion. We measure overlaps in a fixed 100-dimensional subspace S, spanned by the leading right singular vectors of within-question residual hidden states on the fitting partition, so that overlaps are taken where responses to the same question differ. With P_S the projection onto S, we write $v_X = P_S W_X / \|P_S W_X\|$. Replacing a training direction u by u' changes the overlap with criterion X by

$$\mathrm{rem}_X = v_X^\top (u - u'), \tag{2}$$

a profile across all criteria that is known before any training (Equation (13)).

For a driver A and a checker B, let

$$a = v_A, \quad b = v_B, \quad s = a^\top b, \quad c = |s|, \qquad u_{A\setminus B} = \frac{a - sb}{\sqrt{1 - c^2}} \quad (c < 1). \tag{3}$$

The control arm trains along $u_A = a$ and the treated arm along $u_{A\setminus B}$, the *blind direction* for B (Figure 2). Among unit directions in S orthogonal to b, $u_{A\setminus B}$ is the one closest to a. It keeps an alignment $\sqrt{1 - c^2}$ with A after a rotation of $\delta = \arcsin c$, which ranges from 21° to 38° for our pairs (Appendix C). The same construction fixes how much each criterion loses. For the checker $\mathrm{rem}_B = s$, so c is exactly the overlap it gives up, whereas the driver gives up only $\mathrm{rem}_A = 1 - \sqrt{1 - c^2}$. This asymmetry is small for moderate overlaps. In pair 1, for example, $c = 0.56$, so the checker's overlap drops from 0.56 to zero while the driver's drops by 0.17 and the treated direction keeps 0.83 of its alignment with A. The orthogonality is exact in coefficient space. Readouts, however, score raw states through $D^{-1}W_X$, and in that map B keeps between 2.7% and 24.1% of its sensitivity to the treated direction. How much of the removal reaches the trained policy is an empirical question.

4 Experiments

4.1 Setup

The main experiments use the Qwen2.5-1.5B base model (Qwen Team, 2024) and FLASK's GPT-4-scored dimensions (Ye et al., 2024), with probe fitting, criterion selection and measurement on three

question-disjoint partitions (FIT2, 534 questions; SELECT and EVAL, 583 each). The five criterion pairs and their orientations were chosen on SELECT for readout validity and response agreement, without reference to c or to any outcome, and frozen before training (Table 1). They are not the pairs most favourable to the construction. Across all 110 ordered pairs of the eleven criteria, the driver keeps more raw-state sensitivity than the checker in 105, and on our pairs the construction removes less of the checker's sensitivity than on a typical pair (Appendix N).

Both arms of a comparison run DPO ($\beta = 0.1$) with LoRA (rank 16) on the same candidate response pairs and hyper-parameters. The candidates are the existing FLASK responses rather than generations, so the two arms see identical text and differ only in which response of each pair is preferred, by u_A in the control arm and by $u_{A\setminus B}$ in the treated arm. Because the two labellings can move the policy at different speeds, we compare arms at equal displacement rather than at equal step counts, choosing for each arm the checkpoint whose conditional token KL from the base model is closest to 0.45 within fixed tolerances. Eight pair–seed cells meet the rules, one for pairs 1 and 5 and two for pairs 2–4 (Appendix D). A single trained arm cannot reveal a coverage gap on its own, because training moves many features at once. In three earlier single-arm runs whose directions overlapped ArmoRM's readout by only 0.04 to 0.06, ArmoRM still detected the trained policy at 0.51 to 0.56 (Appendix M). The two-arm comparison removes this shared response and isolates the effect of changing the direction. The primary quantity is the loss of checker detection and its overlap-weighted aggregate,

$$\Delta_{B,j} = \alpha^{\mathrm{ctrl}}_{B,j} - \alpha^{\mathrm{treat}}_{B,j}, \qquad T = \frac{\sum_j c_j \Delta_{B,j}}{\sum_j c_j}, \tag{4}$$

estimated over the 61,159 EVAL response pairs with seeds averaged within pairs and a question-clustered bootstrap. We also express the driver's retention and the checker's loss as fractions of the control arm's detection above chance, $R_A = (\alpha^{\mathrm{treat}}_A - q_A)/(\alpha^{\mathrm{ctrl}}_A - q_A)$ and $E_B = (\alpha^{\mathrm{ctrl}}_B - \alpha^{\mathrm{treat}}_B)/(\alpha^{\mathrm{ctrl}}_B - q_B)$.

4.2 Coverage under ordinary reward optimisation

Before constructing anything, we ask how the criteria respond to an ordinary training direction. We train along the projection into S of a readout fitted to ArmoRM's helpfulness, correctness and coherence scores (Wang et al., 2024a), and compare the result at matched KL with an arm trained along a random direction in S that overlaps little with every criterion (Appendix W). The reward model's own attribute heads detect the reward arm above chance (0.534 to 0.545), as they should.

For each FLASK criterion we compare how much more it detects the reward arm than the random arm with its overlap with the reward direction. Ten criteria register the reward arm clearly. `harmlessness`, whose readout is nearly orthogonal to the reward direction, shows no detectable gain. Across the eleven criteria the gain follows the overlap (Spearman $\rho = +0.71$), although among the ten criteria with substantial overlap the order is loose. The same holds for the ambient ArmoRM-derived direction used in Section 4.4. In its two control runs, the overlap of each criterion with the direction orders how well that criterion detects the trained policy (Spearman $\rho = +0.93$ over 22 run–criterion observations; Appendix F). Nothing was removed from either direction, so the unevenness comes from ordinary reward optimisation, and a checker chosen for its agreement with the reward model on responses can still register little of what the reward model trains.

4.3 Removing the checker's share

Table 2 compares the two arms of each pair. In every pair the treated arm is detected less by its checker, and the aggregate loss is $T = +0.0633$ $[+0.0596, +0.0670]$; leaving out any one pair keeps it between $+0.056$ and $+0.068$. The driver loses much less. It keeps between 79% and 99% of its detection above chance, while the checker loses between a quarter and just over half of its own. In pair 2 the contrast is plainest. `logical efficiency` detects the two arms almost equally (0.645 and 0.644), whereas `conciseness` falls from 0.539 to 0.500, close to chance. In the other four pairs the checker still detects the treated arm well above chance. The loss is also not an artefact of weak checkers, since in the control arms every checker detects the policy change at 0.54 to 0.74. Across all fifty-five readout contrasts, the checker's is the largest in each pair (Figure 3). Figure 1 (right) shows one comparison in full. In pair 1 the checker `insightfulness` loses 0.089 of detection, nearly twice as much as any other readout, the driver `completeness` loses 0.048, and the remaining nine criteria lose between 0.026 and 0.044, in rough order of the overlap each of them gives up.

Table 2: Main result. Detection rates α are seed-averaged and Δ_B is the between-arm difference on the checker with a question-clustered interval; R_A and E_B use the same rates above the empirical chance reference $q_X \simeq 0.468$ (Appendix G). Overlaps c are in Table 1.

#	driver $A \to$ checker B	$\alpha_A^{\text{ctrl}} \to \alpha_A^{\text{treat}}$	$\alpha_B^{\text{ctrl}} \to \alpha_B^{\text{treat}}$	Δ_B	95% CI	R_A	E_B
1	completeness → insightfulness	.7684 → .7204	.7395 → .6506	+.0889	[+.0809, +.0962]	84.0%	32.7%
2	logical efficiency → conciseness	.6454 → .6440	.5391 → .5004	+.0387	[+.0332, +.0445]	99.2%	54.2%
3	logical robustness → comprehension	.6812 → .6369	.6893 → .6139	+.0754	[+.0683, +.0824]	79.2%	34.1%
4	logical correctness → commonsense	.6444 → .6265	.6688 → .6082	+.0606	[+.0521, +.0688]	89.8%	30.2%
5	readability → factuality	.7085 → .7010	.6242 → .5859	+.0383	[+.0336, +.0429]	96.9%	24.5%
contrast-weighted aggregate T				**+.0633**	[+.0596, +.0670]		
unweighted mean $\bar{\Delta}_B$				+.0604	[+.0569, +.0639]		

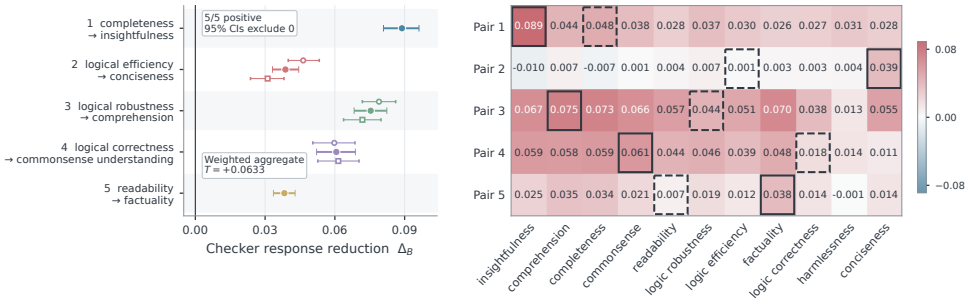

Figure 3: **The checker loses the most detection in every pair.** Left: loss of checker detection Δ_B for each pair (filled circles) with 95% question-bootstrap intervals; hollow circles and squares are seeds 0 and 1 where two seeds are available. Right: detection lost by all eleven readouts in each pair; solid borders mark the checker B, dashed borders the driver A. The weighted aggregate is $T = +0.0633$.

The treated arm differs from the control in two respects. A component along b was removed, and some component of that length was removed at all. All removals of length c from a form a cone of half-angle δ around a, and $u_{A\setminus B}$ is the one generator of that cone that removes B's component. To separate the two effects we train placebo arms along another generator, $\text{norm}(a - c\,r)$ for a random unit $r \in S$ with $a^\top r = c$. Such a direction removes a component of the same length, has the same residual norm and makes the same angle with a as $u_{A\setminus B}$, but points elsewhere (Appendix V). Three pairs yield a KL-matched placebo arm. In each of them the directional removal costs the checker more than the placebo does, by $+0.080$, $+0.036$ and $+0.047$, all with intervals above zero (Figure 6a). The placebo itself is not inert, and in pair 3 it accounts for more than a third of the effect.

4.4 Overlap orders every criterion

The construction changes the overlap with all ten non-driving criteria, not only with the checker, and rem_X gives the size of each change before training. Figure 4a places the fifty measured losses against these predictions. The two orders agree closely (pooled Spearman $\rho = +0.81$), and within every pair the checker, whose overlap change is the largest by construction, sits at the top. The agreement holds at both ends. In pair 2, `insightfulness` and `completeness` lose almost no overlap and are in fact detected slightly more in the treated arm, and `harmlessness`, which has the smallest predicted change in pairs 3–5, has the smallest or second-smallest measured loss there. The drivers lose overlap $1 - \sqrt{1 - c^2}$, between 0.07 and 0.22, which places them among the other criteria, so their small loss in Table 2 is itself anticipated. A family of six further runs that rotate a driving direction away from a group of checked criteria by 15° to 60° shows the same ordering at a fixed step ($\rho = +0.74$ over 66 run–criterion observations), although those runs are not matched on KL (Appendix E). The ordering is weakest in pair 2 ($\rho = +0.22$; Figure 4b), where the removed component is smallest ($c = 0.36$) and the predicted changes are compressed into a narrow range (Appendix N).

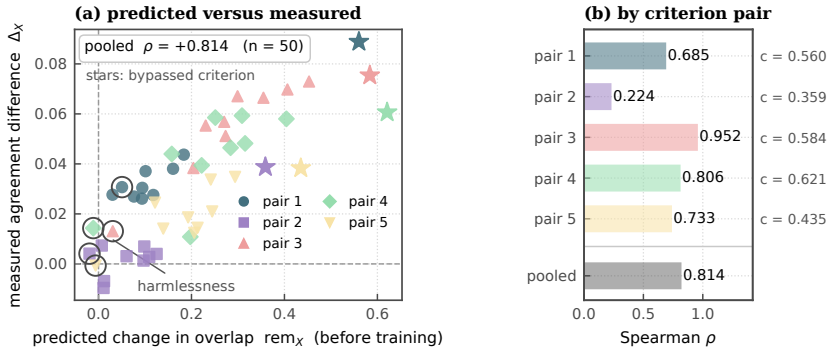

Figure 4: **Overlap computed before training orders what each criterion loses.** (a) Detection lost by each of the fifty non-driving readouts in the five constructed comparisons (the agreement difference Δ_X) against the change in overlap rem_X computed before training. Stars mark the checker B (the bypassed criterion) of each pair and circles `harmlessness`. (b) Spearman correlation between rem_X and Δ_X within each pair and pooled, with the removed component c of each pair.

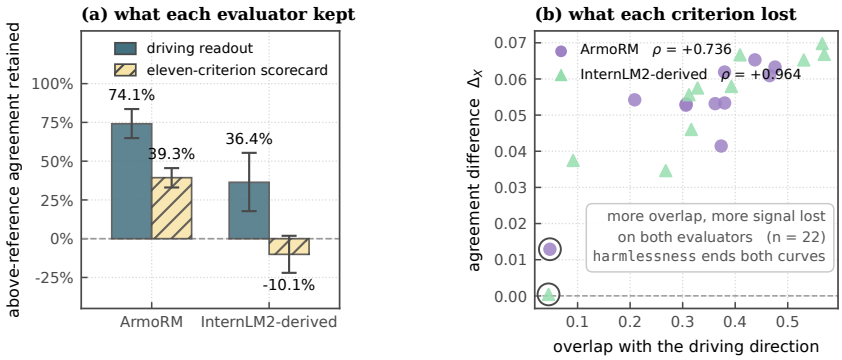

Figure 5: **External reward-derived signals.** (a) Share of detection above the fixed 0.5 reference that the treated arm keeps, for the driving readout and for the mean of the eleven FLASK readouts. This descriptive summary differs from the chance-adjusted ratios of Table 2, and the displayed ranges vary only the treated arm. (b) Detection lost by each FLASK criterion against its overlap with the driving direction, computed before training. `harmlessness`, ringed, has the smallest overlap and the smallest loss under both signals. Detection is length-adjusted, with question-clustered bootstrap intervals (Appendix F).

The same pattern appears when the training signal comes from outside the criterion set. We take a direction fitted to ArmoRM's attribute scores as the driver and remove from it a separately fitted HelpSteer2 attribute subspace (Wang et al., 2024b). The driver's own readout still detects the treated arm above chance, while all eleven FLASK criteria, which play no part in the construction, lose detection roughly in the order of their overlap with the driving direction (Figure 5b, $\rho = +0.74$), and more tightly ($\rho = +0.96$) when the direction comes from InternLM2-1.8B-Reward (Cai et al., 2024) with the joint span of the eleven readouts removed. Measured against the fixed 0.5 reference, the driving readout keeps 74% and 36% of its detection under the two signals, while the mean of the eleven criteria keeps 39% and falls below the reference (Figure 5a).

4.5 Robustness and Timing

Figure 6a collects the variations of the main comparison on a single axis. Both seeds of pairs 2–4 show the loss, although the two seeds of pair 2 differ by more than their question-level intervals allow (0.046 and 0.031). The aggregate changes little when readouts are adjusted for response length or when only the 112 questions outside the DPO candidate pool are scored. Refitting everything on Qwen2.5-0.5B gives the same aggregate on the four pairs that admit a matched cell. On Qwen2.5-7B,

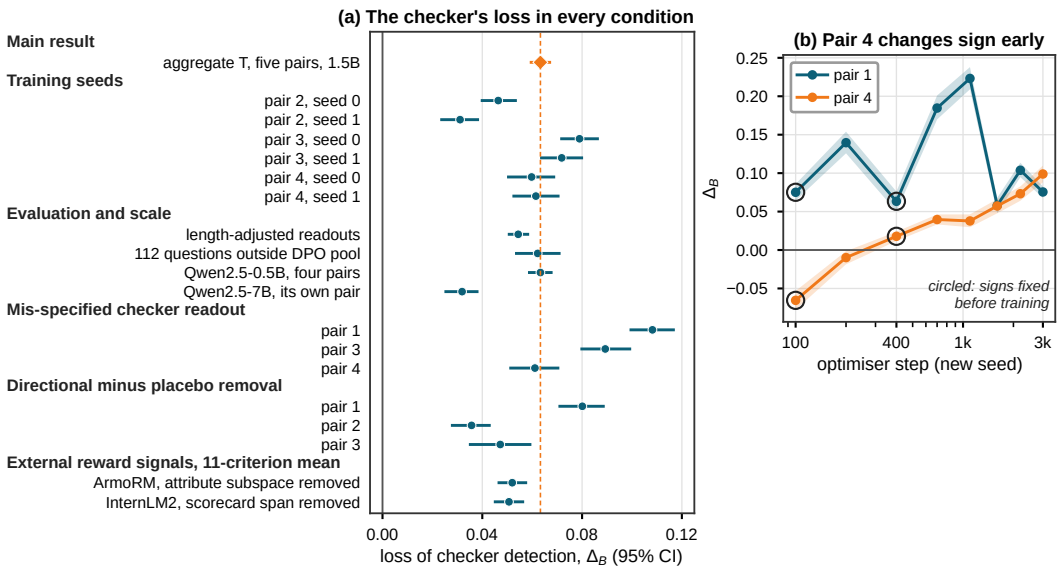

Figure 6: **The checker's loss across conditions and over training.** (a) Loss of checker detection with 95% question-bootstrap intervals for the main aggregate (orange, dashed reference) and for each variation: individual training seeds, length-adjusted readouts, questions outside the DPO candidate pool, two further model scales, a checker readout refit on half the data with a ten-fold ridge penalty, the directional minus placebo contrast, and the eleven-criterion mean under two external reward-derived signals. (b) Checker loss over optimiser steps on an independently trained seed; circled points are the steps whose signs were fixed before training, and shaded bands are pointwise intervals. Appendices H, K, L, V, F and I.

Table 3: First-order prediction against measurement at 100 optimiser steps; predictions were hashed before the measured column existed. The matched-KL column is the main result, shown because the two regimes need not agree.

pair	predicted	measured at step 100	95% CI	matched-KL
1	+0.0571	+0.0788	[+0.0667, +0.0904]	+0.0889
2	+0.0237	+0.0253	[+0.0187, +0.0324]	+0.0387
3	+0.0247	+0.0029	[−0.0026, +0.0082]	+0.0754
4	**−0.0848**	**−0.0904**	[−0.1049, −0.0764]	+0.0606
5	+0.0363	+0.0076	[+0.0013, +0.0141]	+0.0383

the same frozen selection rule picks a different pair, which shows a smaller but clear loss, with the driver keeping 94% of its detection (Appendix L). Rebuilding the treated direction from a deliberately mis-specified checker readout and measuring with the true one leaves the loss intact in all three pairs tested, so the construction needs white-box access to an approximate readout rather than to the exact one. Under the two external reward-derived signals of Section 4.4, the mean detection of the eleven criteria falls by about 0.05 in both cases.

The size of the loss also depends on when it is measured. Before training, we accumulated the preference gradient over the first 100 batches at the base model, scaled it by bisection on KL alone to match each arm's 100-step checkpoint, evaluated the displaced model with the unchanged measurement pipeline, and recorded the predicted checker contrast before any checkpoint was measured. The prediction had the right sign in all five pairs (Table 3), including pair 4, whose checker at step 100 detects the treated arm more than the control. On an independently trained seed, pair 4 starts negative and crosses zero before step 400, while pair 1 stays positive throughout (Figure 6b). Two further predictions made in advance, the ordering of criteria within each pair and the sign of the placebo contrast at step 100, were not met (Appendix Y). Replacing the update direction at fixed states gives the same picture. For pair 4 at step 100, the preconditioned form of momentum

raises checker detection by about 0.07 in both arms, but hardly changes it at step 400 (Figure 10). A checker that registers training at one stage can miss it at another.

4.6 When the construction is available

The construction needs two readouts that share a removable component without sharing so much that removing it erases the driver. Across all 110 ordered pairs of the eleven criteria this is common. The median overlap is $c = 0.40$, and the checker keeps at most a quarter of its raw-state sensitivity in 89 pairs (Appendix N). Our five pairs have larger overlaps than most (median percentile 80%) and leave the checker more of its sensitivity than a typical pair (median percentile 71%), so they do not favour the construction.

Whether two criteria admit such a direction also depends on how their readouts are fitted. In an exploratory path over the ridge penalty, run on readouts fitted to ArmoRM's attribute heads, a stronger penalty lowers the effective dimensionality of the readouts and raises their overlap until no cell admits a usable construction; whitening the features removes this dependence (Appendix R). The path does not enter any estimate above.

5 Discussion

Two features of the setting keep the loss partial. Readouts score raw states, where the treated direction keeps up to a quarter of the checker's sensitivity, and DPO moves the policy along accumulated preference gradients rather than along $u_{A \setminus B}$ itself, as the early reversal in pair 4 shows. Neither c nor the residual sensitivity predicts Δ_B across pairs, and a confirmation round refuted the prediction that a pair with smaller residual sensitivity would lose more (Appendix N.1). What overlap supports is the ordering of criteria within a comparison.

The evidence also has a limited scope. On generated text, an independent LLM judge separates the two arms of pairs 1 and 3 on the checker criterion by 0.33 and 0.62 rubric points but moves the driver criterion nearly as much (Appendices T and U), so a checker-specific loss on generated text remains to be shown. Nonlinear readouts did not pass the validity gate, IPO and SimPO (Azar et al., 2024; Meng et al., 2024) gave no admissible KL-matched comparison (Appendix J), and the reversed orientation of pair 3 never reached the KL target (Appendix P). Keeping only pairs whose GPT-4 scores agree better with human ratings leaves T between $+0.061$ and $+0.075$ (Appendix Q). Our intervals hold the trained checkpoints fixed, so the aggregate describes a consistent effect over five pairs rather than an estimate over training runs (Appendix H).

A direction built against one checker stays visible to others. The criterion that agrees most with the driver on responses still detects the treated arm at 0.60 to 0.67 (Appendix X). Which checkers to add is harder to say, since no rule we tried, whether by agreement, by geometric sensitivity or by the overlap lost to the intervention, clearly beats random choice. A criterion trusted to monitor a training direction should therefore have its overlap with that direction and its detection of an unmodified control run measured before training, and its detection followed against a matched control at several points of training. With several checkers, the coverage of their joint span matters more than their number (Appendix O), and averaging validated readouts into a scorecard can by itself reduce `conciseness` to chance (Appendix S).

6 Conclusion

Criteria that rank responses alike can register very different amounts of the same training change. For linear readouts on a shared representation, that amount is fixed to first order by an overlap computed before training, and the overlap ordered the criteria's responses under ordinary reward optimisation, a direction that removes a checker's share, and external reward signals. Removing that share lowered the checker's detection in every pair while the driver kept most of its signal. A criterion chosen to monitor optimisation should be judged by its coverage of the training direction as well as by its agreement on responses.

AI USE STATEMENT

In this work, we used generative AI tools to help develop the conceptual framework, to implement methods and to assist with translation. We have not used generative AI tools to formulate mathematical claims, to provide ingredients for or write proofs, to propose or refine hypotheses, to design or give feedback on the research methodology or experiments, to clean or reformat datasets, or to interpret results. Generating synthetic data sets and qualitative or thematic data analysis are not applicable to this work; the candidate responses are those released with FLASK. Additionally, we used generative AI tools to create and edit software code, to create and modify scientific figures, to draft parts of the paper, to edit it for readability, to suggest its structure and to propose its title. The GPT-4 scores of FLASK and the LLM judge of Appendix T are measurement instruments of the experiments and are described with them. We have reviewed all AI-assisted work. Code was tested and its outputs checked against the experimental logs, every reported number was traced to the analysis scripts and logs, and all text, claims and figures were checked by the authors. The authors take full responsibility for the content of this paper, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work characterises a failure mode of multi-criteria evaluation. The construction requires white-box access to a model's representations and to the readouts of the criteria being bypassed. We use it as a measurement instrument for auditing directional coverage, so that evaluation protocols can check what their criteria cover.

REPRODUCIBILITY STATEMENT

Partitions, criterion pairs, orientations and acceptance criteria were frozen before training, with all inputs pinned by SHA-256. Geometric identities are checked per pair at construction time. The main KL acceptance rules, bootstrap procedure and treatment of unmatched cells are specified in Appendix D; the 7B checkpoint-selector difference is specified in Appendix L. Pre-specified hypotheses, acceptance criteria and decision rules for the experiments reported in the paper are documented in the appendices, with confirmatory and post-hoc analyses distinguished (Appendix B). The corresponding frozen protocols with their hashes, the criterion readouts and directions, the per-checkpoint measurements and the bootstrap draws underlying every table and figure are released with this submission; the training and analysis code will be released upon publication.

REFERENCES

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi 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 (AISTATS)*, 2024. arXiv:2310.12036.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Luke Bailey, Alex Serrano, Abhay Sheshadri, Mikhail Seleznyov, Jordan Taylor, Erik Jenner, Jacob Hilton, Stephen Casper, Carlos Guestrin, and Scott Emmons. Obfuscated activations bypass LLM latent-space defenses. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2412.09565.

Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y. Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. InternLM2 technical report. *arXiv preprint arXiv:2403.17297*, 2024.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.02743.

Sumanth Doddapaneni, Mohammed Safi Ur Rahman Khan, Sshubam Verma, and Mitesh M. Khapra. Finding blind spots in evaluator LLMs with interpretable checklists. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 16279–16309, 2024.

Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alex D'Amour, DJ Dvijotham, Adam Fisch, Katherine Heller, Stephen Pfohl, Deepak Ramachandran, Peter Shaw, and Jonathan Berant. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2312.09244.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866, 2023. arXiv:2210.10760.

Jia Liu, Jiaxin Luo, Xinhao Qiu, Yixue Hao, and Min Chen. The geometry of narrow fine-tuning degradation: Trajectory lock-in and spectral bifurcation. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, volume 306 of *PMLR*, 2026. To appear.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2511–2522, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153.

Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2405.14734.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7237–7256, 2020. doi: 10.18653/v1/2020.acl-main.647.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10582–10592, 2024a. doi: 10.18653/v1/2024.findings-emnlp.620.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. HelpSteer2: Open-source dataset for training top-performing reward models. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024b. arXiv:2406.08673.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. FLASK: Fine-grained language model evaluation based on alignment skill sets. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2307.10928.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023. arXiv:2306.05685.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

APPENDIX CONTENTS

	Appendix	Page
A	Notation and glossary	12
B	Pre-specified tests and scope	14
C	Construction geometry and scoring coordinates	14
D	Experimental protocol and agreement statistic	16
E	Controlled-angle construction and analysis	18
F	External reward-derived subspace experiment	20
G	The chance level of the agreement statistic	21
H	Robustness diagnostics	22
I	Independent-seed replication of the trajectory prediction	23
J	Readout and optimisation scope	24
K	A mis-specified checker readout	25
L	A third model scale: Qwen2.5-7B	26
M	Single-arm response and the two-arm estimand	27
N	How common is the construction? A census over all ordered pairs	28
O	Choosing checkers: a coverage audit	30
P	Reversed orientation: a control that yielded no usable cell	31
Q	Robustness across human–model agreement thresholds	32
R	When does the construction exist? A regularisation path	32
S	Per-criterion validity does not imply aggregate validity	33
T	An independent judge scoring newly generated text	35
U	An external reward model scoring newly generated text	36
V	Equal-length placebo arms	37
W	Ordinary reward-model optimisation: per-criterion detail	39
X	Checker selection: rankings and per-pair cells	40
Y	First-order mechanism: criteria, traces, and provenance	41

A NOTATION AND GLOSSARY

Symbols are listed in the order they are introduced, with the location of their definition in the body; the terms of the paper's vocabulary follow.

symbol	first defined	meaning
Representations and readouts		
h	§3	A response representation (hidden state), $h \in \mathbb{R}^d$
μ, σ, D	§3; App. C	Fitting-set mean and per-coordinate scales; $D = \mathrm{diag}(\sigma)$
W_X	§3	Fitted coefficient vector of criterion X in standardised coordinates
$s_X(h)$	§3	Readout score of criterion X, $s_X(h) = [D^{-1}(h-\mu)]^\top W_X$
U, S, P_S	App. C	Measurement basis $U \in \mathbb{R}^{d\times 100}$, its span S, and $P_S = UU^\top$
The construction		
a	§3	Driving direction, $a = P_S W_A / \|P_S W_A\|$; the control arm trains along $u_A = a$
b, $\tilde{b}$	§3	Checker direction and its sign-aligned form $\tilde{b} = \mathrm{sign}(s)\, b$
s, c	§3	Signed overlap $s = a^\top b$ and directional overlap $c = \|s\| = a^\top \tilde{b}$
$u_{A\setminus B}$	§3	Blind direction, $(a - sb)/\sqrt{1-c^2} = \mathrm{norm}(a - c\,\tilde{b})$; orthogonal to b
δ	§3	Rotation from a to $u_{A\setminus B}$, $\delta = \arcsin c$

symbol	first defined	meaning (continued)
v_X	§3	Projected direction of any criterion X, $v_X = P_S W_X / \|P_S W_X\|$
rem_X	§3	Predicted change in coefficient overlap, $v_X^\top (u_A - u_{A\setminus B})$
η_X, η_A	App. C	Residual scoring sensitivity $\|u^\top (W_X/\sigma)\| / \|a^\top (W_X/\sigma)\|$; η_A is the same for the driving criterion
$R(c)$	§4.3	Equal-angle set $\{r \in S : \|r\| = 1,\ a^\top r = c\}$; $\tilde{b}$ and the placebo direction are both members
C, P_C, u_C	App. O	A checker set, the projection onto its projected-coefficient span, and the direction orthogonal to that span
u_{RM}	§4.4	Ambient readout direction fitted from an external reward model's attribute scores
u_{RM}^S	App. W	Its projection into the measurement subspace, $\mathrm{norm}(P_S u_{\mathrm{RM}})$; the direction the reward arm trains along
The endpoint		
q_i, y_{i1}, y_{i2}	App. D	A question and its two responses; $N = 61{,}159$ such pairs
m_t, ℓ_t	App. D	Mean response-token log-probability at step t, and its change from the base model
$d_{t,i}$	App. D	Policy difference within pair i, $\ell_t(q_i, y_{i1}) - \ell_t(q_i, y_{i2})$
$z_{X,i}$	App. D	Criterion difference within pair i, $s_X(h_{i1}) - s_X(h_{i2})$
$A(d, z)$	App. D	Agreement credit: $\frac{1}{2}$ if $d = 0$, 1 if $dz > 0$, else 0
α_X	App. D	Agreement of criterion X, $N^{-1} \sum_i A(d_{t,i}, z_{X,i})$
q_X	§4.1	Empirical chance reference for α_X, $q_X \simeq 0.468$
$\mathcal{K}_t$	App. D	Conditional token KL from the base model; arms are matched at $\mathcal{K}^* = 0.45$
Contrasts and aggregates		
$\Delta_{B,j}$	§4.1	Between-arm contrast on the bypassed criterion, $\alpha_{B,j}^{\mathrm{ctrl}} - \alpha_{B,j}^{\mathrm{treat}}$
T, c_j	§4.1	Contrast-weighted aggregate $\sum_j c_j \Delta_{B,j} / \sum_j c_j$ with weights $c_j = \|a_j^\top b_j\|$; $J = 5$ pairs
$\bar{\Delta}_B$	§4.1	Unweighted mean of $\Delta_{B,j}$ over the same pairs
R_A, E_B	§4.1	Driving-signal retention and bypassed-signal removal, both above q_X
Checker selection		
$C_2(A, X)$	App. X	Share of same-question pairs on which two readouts differ in the same direction
$S_R(j, k)$	App. X	Top-k criteria chosen by rule R for pair j
M_R, V_R	App. X	Maximum retained detection and count above the detection threshold over $S_R(j, k)$

The vocabulary is $\mathcal{V}$ (Appendix D); V denotes a criterion group in Appendices E and F. The whitening exponent γ is defined in Appendix R. The recurring terms of the paper's vocabulary are collected below.

term	meaning
agreement–coverage gap	The discrepancy this paper studies: two criteria can rank the same responses alike while covering very different amounts of a given optimisation direction.
detection signal	Agreement α_X above the stated reference. Main-table ratios use $q_X \simeq 0.468$; secondary summaries retaining the fixed 0.5 reference are labelled explicitly.
c, directional overlap	$a^\top \tilde{b}$ between the two criteria's normalised projected coefficient directions. Computed on the base model.
$\delta = \arcsin c$	The rotation from a to the blind direction $u_{A\setminus B}$.
η_X, residual scoring sensitivity	$\|u^\top (W_X/\sigma)\| / \|a^\top (W_X/\sigma)\|$: how much of criterion X's raw-state scoring sensitivity the intervention leaves. Computed on the base model.
blind direction $u_{A\setminus B}$	$\mathrm{norm}(a - c\tilde{b})$: the driving direction with the component shared with B removed.
equal-angle set $R(c)$	$\{r \in S : \|r\| = 1,\ a^\top r = c\}$. $\tilde{b}$ is one member; the placebo direction is another.
Δ_B	$\alpha_B^{\mathrm{ctrl}} - \alpha_B^{\mathrm{treat}}$ at matched KL: the between-arm difference in the bypassed criterion's agreement.

term	meaning (continued)
T	The contrast-weighted aggregate of Δ_B over the five pairs; the primary inferential quantity.
R_A, E_B	Retention of the driving signal and removal of the bypassed signal, both relative to the control arm's above-reference agreement.
matched KL	The comparison protocol: checkpoints of the two arms selected so their conditional token KL from the base model matches a fixed target under fixed tolerances, independently of any agreement statistic.

B PRE-SPECIFIED TESTS AND SCOPE

For the confirmatory experiments reported here, hypotheses, acceptance gates and decision rules were fixed before the corresponding outcomes were inspected; input hashes and any dated amendments accompany the protocols. Post-hoc analyses are identified where reported. The main decision rules and outcomes appear in Appendix D; the placebo, census-confirmation, first-order and independent-seed tests retain their own criteria and outcomes in Appendices V, N.1, Y and I. A supported component is distinguished from a conjunction that was not met; neither the original prediction nor its status is redefined after observing the result.

Scope of the mechanism tests. The controlled evidence establishes directional detection separation for linear criterion readouts under DPO. Pre-specified nonlinear-readout extensions did not yield criterion pairs passing the required validity gates for an interpretable paired test. Nonlinear-readout generality and transfer to other preference objectives are therefore outside the claims established here (Appendix J).

C CONSTRUCTION GEOMETRY AND SCORING COORDINATES

What the instrument does. We construct a direction orthogonal to one projected probe coefficient while retaining overlap with another. The experiments measure how using this direction to label DPO pairs changes agreement with fixed evaluation readouts. Coefficient geometry and scoring sensitivity are specified separately below.

Equal-length removals lie on a cone. Let r be any unit vector in S with $a^\top r = c$, and remove a component of length c from a along it:

$$u = \frac{a - cr}{\sqrt{1 - c^2}}, \qquad a^\top u = \sqrt{1 - c^2}, \tag{5}$$

so every such direction makes the same angle $\delta = \arcsin c$ with a. The equal-length removals therefore form a cone of half-angle δ about a, and $u_{A\setminus B}$ is one of its generators, the one obtained by taking $r = \tilde{b}$, with $a^\top \tilde{b} = c$.

We use this only to say what a control can hold fixed. A second generator of the same cone removes the same length, leaves the same residual norm $\sqrt{1 - c^2}$, and rotates a by the same angle; it differs from $u_{A\setminus B}$ only in *which* direction was removed. Two statements that do not follow are worth marking explicitly. A criterion X whose projected direction v_X satisfies $a^\top v_X \neq c$ need not correspond to any single generator of this cone, so the cone is not a catalogue of one blind direction per criterion. And the cone's half-angle depends only on c, not on how many criteria are used as checkers; this says nothing about whether additional checkers cover these directions. A direction orthogonal to one criterion may be visible to another, and evading several criteria at once concerns their joint orthogonal complement, which Appendix O treats directly.

Inputs, outputs, access. The inputs are a frozen language model, scored responses, and a fitting split for criterion probes. The construction returns $c \in [0, 1)$ and a unit direction $u_{A\setminus B}$. It uses hidden states and fitted coefficients, without gradients through the judge that produced the scores. Probe fitting, criterion selection, and outcome measurement use question-disjoint FIT2, SELECT, and EVAL partitions. Training-candidate exposure and evaluation-set length adjustment are specified separately.

Procedure. Four steps, each taken before any training.

1. **Freeze a representation.** Let $h \in \mathbb{R}^d$ be the mean hidden state over response tokens, excluding the prompt, with input sequences truncated at 512 tokens. On FIT2, compute $\mu = \mathrm{mean}(h)$ and $\sigma_k = \mathrm{std}(h_k) + 10^{-6}$, retaining all coordinates. Write $D = \mathrm{diag}(\sigma)$ and $x = D^{-1}(h - \mu)$.
2. **Fit a probe per criterion.** Ridge-regress the centred criterion scores on x, without an intercept or target standardisation:
$$W_X = (X^\top X + \lambda n_f I)^{-1} X^\top (y_X - \bar{y}_X \mathbf{1}), \qquad \lambda = 1. \tag{6}$$
Here X contains the standardised FIT2 feature rows with finite target scores, n_f is their count, and I is the identity matrix. The scalar readout is $s_X(h) = x^\top W_X$; a constant target offset does not affect pair rankings.
3. **Choose a measurement basis.** Compute within-question residuals of the *unstandardised* centred FIT2 hidden states. The leading 100 right singular vectors of this 8010×1536 residual matrix form $U \in \mathbb{R}^{1536\times 100}$. Define $S = \mathrm{range}(U)$, $P_S = UU^\top$, and $P_S^\perp = I - P_S$. The implemented construction applies this projection directly to the standardised-space coefficients W_X. All vector norms and geometric overlaps below use the Euclidean inner product. We write $\mathrm{norm}(v) = v/\|v\|$ only for $v \neq 0$, and $W_X/\sigma = D^{-1}W_X$ denotes coordinate-wise division. The sign-alignment convention is $\mathrm{sign}(0) = 1$, so $\tilde{b}$ remains a unit vector also when $s = 0$. This defines the implemented coefficient-space intervention; the raw-state scoring map is evaluated separately through $D^{-1}W_X$.
4. **Build the direction.** For the pair (A, B), normalise both readouts inside S, record the overlap c, and remove it.

Step 4 applies to readouts with $\|P_S W_A\| > 0$ and $\|P_S W_B\| > 0$. Define

$$a = \frac{P_S W_A}{\|P_S W_A\|}, \qquad b = \frac{P_S W_B}{\|P_S W_B\|}, \qquad s = a^\top b, \qquad c = |s|, \qquad u_A = a, \tag{7}$$

and, for $c < 1$, the *blind direction*

$$u_{A\setminus B} = \frac{a - s\,b}{\|a - s\,b\|}. \tag{8}$$

Let $P_B = bb^\top$ be the orthogonal projection onto $\mathrm{span}\{b\}$. The signed overlap is $s = a^\top b$, while $c = |s|$ records its magnitude. Define $\phi = \arccos c \in [0, \pi/2]$ as the smaller angle between the readout axes and $\delta = \angle(u_A, u_{A\setminus B})$ as the rotation of the training direction. Then $c = \cos\phi$, $\delta = \arcsin c$, and $a^\top u_{A\setminus B} = \sqrt{1 - c^2}$. Here c is a geometric overlap of readout directions; it is distinct from an empirical correlation between criterion scores.

Exact geometric identities. The construction gives

$$\begin{aligned} \|P_B u_A\|^2 &= c^2, \\ \angle(u_A, u_{A\setminus B}) &= \arcsin c, \\ b^\top u_A - b^\top u_{A\setminus B} &= s, \qquad \left|b^\top u_A - b^\top u_{A\setminus B}\right| = c. \end{aligned} \tag{9}$$

The signed component removed from the unit readout b is s, with magnitude c. Conditional on the fitted readouts, these identities are exact and known before training; the resulting change in agreement is an empirical endpoint.

Coefficient orthogonality and scoring sensitivity. Since $u_{A\setminus B} \in S$, the implemented construction therefore satisfies

$$W_B^\top u_{A\setminus B} = (P_S W_B)^\top u_{A\setminus B} = 0. \tag{10}$$

This is exact orthogonality to the coefficient vector in the construction's Euclidean metric. The scoring map has a different directional derivative in raw hidden-state coordinates. With $g_X = D^{-1}W_X$,

$$s_X(h + \epsilon u) - s_X(h) = \epsilon g_X^\top u. \tag{11}$$

Table 7: Predicted overlap change versus measured agreement difference, by criterion pair. ρ is Spearman's rank correlation between rem_X and Δ_X over the ten non-driving criteria.

#	$A \to B$	c	range of rem_X	ρ
1	completeness → insightfulness	0.560	$[+0.030, +0.560]$	$+0.685$
2	logical efficiency → conciseness	0.359	$[-0.019, +0.359]$	$+0.224$
3	logical robustness → comprehension	0.584	$[+0.031, +0.584]$	$+0.952$
4	logical correctness → commonsense und.	0.621	$[-0.011, +0.621]$	$+0.806$
5	readability → factuality	0.435	$[-0.006, +0.435]$	$+0.733$
pooled ($n = 50$)				$\mathbf{+0.814}$

Thus the coefficient identity does not impose $g_B^\top u_{A\setminus B} = 0$. The measured ratio of treated to control scoring sensitivity,

$$\eta_{B,j} = \frac{|g_{B_j}^\top u_{A_j \setminus B_j}|}{|g_{B_j}^\top u_{A_j}|}, \tag{12}$$

is defined when $|g_{B_j}^\top u_{A_j}| > 0$, as in all five main pairs. It is 18.1%, 24.1%, 14.6%, 19.7%, and 2.7% for pairs 1–5. The construction therefore strongly reduces this scoring sensitivity while retaining a nonzero residual. Our empirical endpoint separately measures the policy preference changes produced by training with the constructed labels.

Where blind directions exist. For $0 < c < 1$, projection removes shared directional overlap while preserving a positive component along A. At $c = 0$, a is already blind to B and remains unchanged. At $c = 1$, the projected coefficient vectors are collinear and no direction within S has zero overlap with b and positive overlap with a. Table 1 reports $c = 0.359$–0.621, corresponding to training-direction rotations $\delta = 21°$–$38°$.

Change in overlap with every other criterion. For a criterion X with nonzero projected readout, define $v_X = P_S W_X / \|P_S W_X\|$. Its signed change in directional overlap is

$$\begin{aligned} \mathrm{rem}_X &= v_X^\top u_A - v_X^\top u_{A\setminus B} \\ &= (v_X^\top u_A)\left(1 - \frac{1}{\sqrt{1-c^2}}\right) + \frac{s\, v_X^\top b}{\sqrt{1-c^2}}. \end{aligned} \tag{13}$$

In particular, $v_B = b$, so $\mathrm{rem}_B = s$ and $|\mathrm{rem}_B| = c$. All five implemented pairs have $s > 0$, without sign flipping, hence $\mathrm{rem}_B = c$ for these pairs. A positive value denotes reduced signed overlap; a negative value denotes increased signed overlap. Equation (13) provides a geometric predictor for the empirical agreement differences tested in Section 4.4.

Table 7 reports this association separately for each criterion pair. In these five pairs, $\mathrm{rem}_B = c$ is the largest observed change among the measured non-driving criteria. The association is least determined in pair 2, which has the smallest c and therefore the narrowest range of predicted changes to order.

D Experimental protocol and agreement statistic

Model, data, partitions. Qwen2.5-1.5B, layer $17/28$, $d = 1536$. Criterion scores come from FLASK (Ye et al., 2024), in which a GPT-4 judge assigns fine-grained per-dimension scores. Questions are split into three disjoint partitions: FIT2 (534 questions, 8,010 responses) for probe fitting, the measurement basis and standardisation; SELECT (583 / 8,742) for criterion-pair selection; and EVAL (583 / 8,741) for every reported measurement. Within EVAL we form all same-question response pairs, 61,159 of them. The effective independent unit is the question, not the pair.

Probe eligibility. On SELECT, a probe is eligible when the question-clustered 95% lower confidence bound of pairwise accuracy exceeds 0.5, the corresponding lower bound of its log-loss

Table 8: The eight pair–seed cells entering the main analysis. Every cell satisfies the relative, absolute and between-arm tolerances.

#	seed	step ctrl	step treat	$\mathcal{K}_{\mathrm{ctrl}}$	$\mathcal{K}_{\mathrm{treat}}$	$\lvert\Delta\mathcal{K}\rvert$
1	0	3000	13000	0.4120	0.4483	0.0363
2	0	1100	3000	0.4171	0.4795	0.0624
2	1	3000	3000	0.5163	0.4350	0.0813
3	0	1600	1600	0.4235	0.4860	0.0625
3	1	1600	1100	0.4508	0.4429	0.0079
4	0	3000	1600	0.3479	0.4282	0.0803
4	1	3000	1600	0.4316	0.4642	0.0326
5	0	4500	4500	0.4303	0.4136	0.0167

improvement over a length-only predictor exceeds zero, and at least 200 pairs are available. Screening uses unequal annotated-score pairs, half credit for probe-prediction ties, and 1000 question-clustered bootstrap resamples. Its length comparison uses length-only versus length-plus-readout logistic predictors. This screening criterion is distinct from the evaluation agreement defined below.

Frozen pairs. Pairs were selected on SELECT and frozen, with all inputs pinned by SHA-256, before any training. Selection uses held-out criterion validity and disjointness; it does not compute c, rem, or any other quantity that predicts the outcome. The orientation rule assigns A to the side with higher held-out validity. Table 1 is the frozen list.

Two arms differing in one direction. We train with DPO (Rafailov et al., 2023) with $\beta = 0.1$ and LoRA ($r = 16$, lr 2×10^{-5}, batch 16, maximum length 512). Candidate responses are not generated: they are the existing FLASK responses to each question, so both arms draw from the same fixed pool and differ only in which member of a pair is labelled preferred. The *control* arm is trained along the unmodified u_A; the *treated* arm along $u_{A \setminus B}$. Candidate response pairs are built with a seed that depends only on the resampling index, independent of the arm; the arm's score decides only which of the *same two* responses is marked preferred. Both arms therefore see identical candidate response pairs, with preferences assigned by their respective training signals. Specifically, the labelling score is $z_u(h) = (h - \mu_0)^\top u$ in raw hidden-state coordinates, with μ_0 the training pipeline's fixed centring vector. This constant does not affect same-question rankings. The higher-scoring candidate is preferred; exact ties select the first sampled candidate. Candidate length filtering is shared across arms, with no arm-specific score filtering.

Matched displacement. The reported KL is the mean conditional forward divergence over response-token contexts in the first 64 entries of the training pipeline's fixed evaluation set EV:

$$\mathcal{K}_t = \frac{1}{N_{\mathrm{tok}}} \sum_{\xi \in \mathcal{T}_{\mathrm{EV}}} \sum_{v \in \mathcal{V}} \pi_t(v \mid \xi) \log \frac{\pi_t(v \mid \xi)}{\pi_0(v \mid \xi)}. \tag{14}$$

Here $\mathcal{T}_{\mathrm{EV}}$ is the multiset of masked response-token contexts, N_{tok} its size, and $\mathcal{V}$ the vocabulary. The vocabulary sum is exact for each context; EV is the pipeline's evaluation set, separate in definition from the FLASK EVAL partition. This is a conditional token-level displacement measure evaluated on fixed contexts.

For each arm, checkpoint selection first minimises $|\mathcal{K}_t - \mathcal{K}^*|$, with $\mathcal{K}^* = 0.45$. The selected checkpoint must satisfy *both* $|\mathcal{K}_t - \mathcal{K}^*|/\mathcal{K}^* \le 0.25$ and $|\mathcal{K}_t - \mathcal{K}^*| \le 0.15$. The two selected arms must also satisfy $|\mathcal{K}_{\mathrm{ctrl}} - \mathcal{K}_{\mathrm{treat}}| \le 0.10$. Unmatched pair–seed cells are excluded. Eight cells enter the main analysis: one seed each for pairs 1 and 5 and two each for pairs 2–4. Table 8 lists them. Because selection targets displacement rather than step, the two arms of a cell often stop at different step counts; five of the eight cells do so.

Why two arms. Coefficient orthogonality characterises the constructed signal; scoring sensitivity and the resulting policy update are measured separately. The empirical test compares two arms sharing candidates and hyperparameters at matched KL displacement. This estimates the difference

induced by changing the training signal under controlled data and displacement conditions. A single-arm illustration appears in Appendix M.

Endpoint. For question q and response y of L_y scored response tokens, define the mean token log-probability and its training change:

$$\begin{aligned} m_t(q,y) &= \frac{1}{L_y}\sum_{k=1}^{L_y}\log \pi_t(y_k \mid q, y_{<k}), \\ \ell_t(q,y) &= m_t(q,y) - m_0(q,y), \\ d_{t,i} &= \ell_t(q_i, y_{i1}) - \ell_t(q_i, y_{i2}), \\ z_{X,i} &= s_X(h_{i1}) - s_X(h_{i2}). \end{aligned} \tag{15}$$

Prompt tokens are excluded from m_t and the hidden states remain frozen. All $N = 61{,}159$ same-question response pairs are included, in the same order on both sides of the comparison. Their agreement credit is

$$A(d,z) = \begin{cases} \frac{1}{2}, & d = 0, \\ 1, & dz > 0, \\ 0, & \text{otherwise.} \end{cases} \qquad \alpha_X = \frac{1}{N}\sum_{i=1}^{N} A(d_{t,i}, z_{X,i}). \tag{16}$$

Equality tests use exact zero, without an epsilon. Policy ties receive half credit, including when both sides tie. Criterion ties with a nonzero policy difference receive zero and remain in the denominator. The point estimate weights response pairs equally, so questions with more pairs have greater weight; uncertainty is clustered by question.

For independent equiprobable, non-tied policy signs, the implemented criterion-tie rule gives a reference $q_X = \frac{1}{2}\Pr(z_X \neq 0)$. The empirical values are approximately 0.468 for the main readouts (Appendix G). Main-table retention and removal use these references. Secondary fixed-0.5 summaries and pre-specified decision thresholds retain their stated definitions.

For frozen criterion pair $j = (A_j, B_j)$, define

$$\begin{aligned} \Delta_{X,j} &= \alpha^{\text{ctrl}}_{X,j} - \alpha^{\text{treat}}_{X,j}, \\ T &= \frac{\sum_{j=1}^{J} c_j \Delta_{B_j,j}}{\sum_{j=1}^{J} c_j}, \qquad c_j = |a_j^\top b_j|. \end{aligned} \tag{17}$$

The main experiment has $J = 5$. We abbreviate the bypassed-criterion contrast as $\Delta_{B,j}$, and as Δ_B when the criterion pair is clear. The unweighted mean is $\bar{\Delta}_B = J^{-1}\sum_{j=1}^{J}\Delta_{B_j,j}$. For subset analyses the sums run over the included criterion pairs. Within each criterion pair, agreement arrays are averaged element-wise over included seeds before computing means and intervals. The same 2000 question-clustered bootstrap index sets are used across arms, seeds, and criterion pairs; the geometric weights c_j remain fixed. The implemented admission rule additionally requires $\alpha^{\text{treat}}_{A,j} \geq 0.70\alpha^{\text{ctrl}}_{A,j}$ for each pair–seed cell. This uses raw agreement, not signal above 0.5; all eight KL-matched cells pass, so this gate excludes none of them.

Descriptive retention and removal. These descriptive ratios were computed after the primary analysis, using seed-averaged agreement values before taking the ratio, and are defined when the control denominator is nonzero. The reference subtracted in both numerator and denominator is the empirical chance level of Appendix G, not 0.5; Table 2 reports $100R_A\%$ and $100E_B\%$ on that basis.

Pre-registered rules and outcomes. Table 9 lists the decision rules fixed before training, together with their outcomes.

E Controlled-angle construction and analysis

This experiment varies a geometric direction within the pipeline's fixed aggregate construction, in which a driving group V and a checked group H of attribute probes replace the single driving and

Table 9: Pre-registered decision rules and outcomes. Rules, tolerances and the stopping condition were fixed before training and all inputs pinned by SHA-256. Every rule is listed, including the one whose predicted ordering was not supported.

rule, fixed before training	outcome	note
Primary inference: lower bound of T's interval above zero	met	$T = +0.0633$, $[+0.0596, +0.0670]$
Direction: $\Delta_B > 0$ in at least four of five pairs	met	five of five
Retention gate: $\alpha_A^{\text{treat}} \geq 0.70\,\alpha_A^{\text{ctrl}}$	met	all eight cells pass; none excluded
KL rule: relative ≤ 0.25, absolute ≤ 0.15, between-arm ≤ 0.10	applied	eight cells matched; unmatched cells left empty
Weights c_j fixed before training	applied	recorded in the frozen geometry file
Stopping rule: five pairs, two arms, two seeds	applied	no additional pairs or seeds were run under this protocol
Predicted ordering: Δ_B increases with c	**not supported**	$\rho_{\text{Spearman}} = +0.500$, $n = 5$, reported as a rank association without a test, as specified

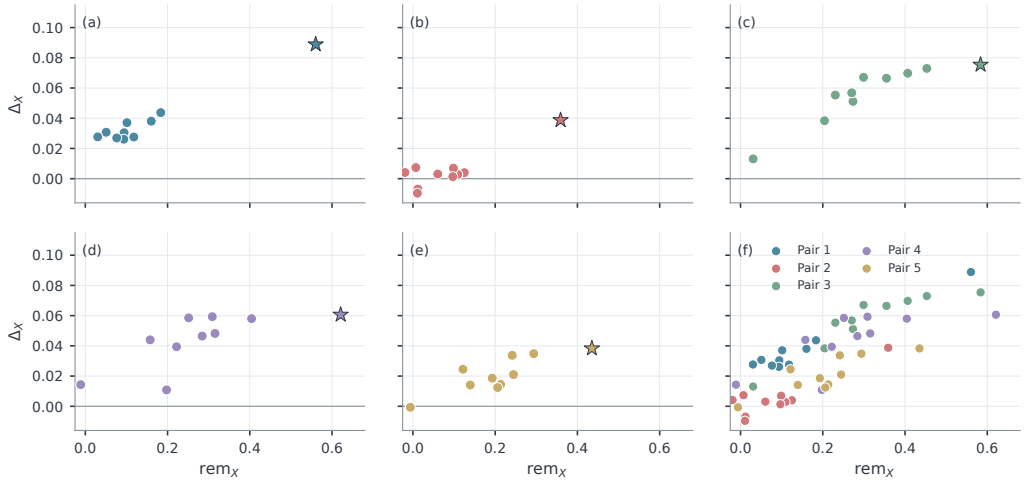

Figure 7: The overlap change computed before training orders the detection differences measured after it. (a) All fifty points: each pair contributes its ten non-driving readouts, the checker drawn as a star; `harmlessness`, ringed, anchors the low-removal end. (b) The association holds within every pair, and is weakest where the construction overlap c is smallest.

checked criteria of the main experiment. Its two starting directions are the normalised means of the coefficient vectors within each group; they are distinct from the five single-criterion pairs in Table 1. Denote their unit vectors by a_* and b_*, with $0 < |a_*^\top b_*| < 1$, and let $P_* = b_* b_*^\top$. Define

$$e_\parallel = \frac{P_* a_*}{\|P_* a_*\|}, \qquad e_\perp = \frac{(I - P_*)a_*}{\|(I - P_*)a_*\|}, \tag{18}$$
$$u_\theta = \cos\theta\, e_\parallel + \sin\theta\, e_\perp, \quad \|P_* u_\theta\|^2 = \cos^2\theta.$$

This identity concerns the construction geometry. We evaluate directions at $15°$, $30°$, $45°$, and $60°$, with seed 0 at all four angles and seed 1 at the two endpoints: six runs in total. The fitting configuration and training hyperparameters are fixed across angles.

Each run is measured at step 3000 for the primary analysis and at step 2000 for the secondary analysis. Thus the twelve run–analysis records represent six training runs, and each pooled analysis contains $6 \times 11 = 66$ run–criterion observations. Geometric overlap is the absolute inner product

Table 10: The six training runs of the rotation experiment. Each run is measured at two steps, so the twelve run–analysis records represent six runs; each pooled analysis contains $6 \times 11 = 66$ run–criterion observations in total.

θ	seed	$\mathcal{K}$ at step 3000	$\mathcal{K}$ at step 2000
15°	0	0.6334	0.3610
15°	1	0.5997	0.3620
30°	0	0.5192	0.3879
45°	0	0.6261	0.5176
60°	0	0.6102	0.5796
60°	1	0.6807	0.5802
pooled $\rho^{\perp}$		$+0.744$	$+0.740$

Table 11: Per-criterion response to the external attribute-subspace intervention. Overlaps are absolute inner products of projected, normalised directions, computed before training. Two control seeds are compared with one treated seed at matched KL; agreement is length-adjusted and intervals are question-clustered bootstrap. `harmlessness` is the only criterion whose overlap increases under the intervention, and it shows the smallest response. The driving readout is listed for comparison in the same units.

readout	overlap before	overlap after	α^{ctrl}	α^{treat}	Δ_X	95% CI
comprehension	0.476	0.057	0.6159	0.5526	$+0.0633$	$[+0.0552, +0.0719]$
completeness	0.466	0.068	0.6336	0.5727	$+0.0609$	$[+0.0530, +0.0687]$
insightfulness	0.438	0.111	0.6309	0.5656	$+0.0653$	$[+0.0573, +0.0735]$
factuality	0.380	0.042	0.5991	0.5457	$+0.0533$	$[+0.0453, +0.0617]$
commonsense understanding	0.380	0.028	0.6068	0.5449	$+0.0619$	$[+0.0539, +0.0700]$
readability	0.373	0.090	0.5733	0.5318	$+0.0414$	$[+0.0334, +0.0491]$
logical robustness	0.362	0.059	0.5763	0.5232	$+0.0532$	$[+0.0447, +0.0620]$
logical efficiency	0.307	0.030	0.5695	0.5165	$+0.0530$	$[+0.0438, +0.0622]$
logical correctness	0.306	0.014	0.5581	0.5053	$+0.0527$	$[+0.0446, +0.0614]$
conciseness	0.209	0.042	0.5684	0.5142	$+0.0542$	$[+0.0456, +0.0631]$
harmlessness	**0.047**	**0.092**	0.5115	0.4986	$+\mathbf{0.0129}$	$[+0.0049, +0.0210]$
eleven criteria, mean			0.5858	0.5337	$+0.0520$	$[+0.0467, +0.0574]$
ArmoRM-derived (driving)			0.6336	0.5990	$+0.0346$	—

of the projected, normalised criterion and intervention vectors. The reported pooled correlations using length-adjusted agreement are $\rho^{\perp} = +0.744$ and $+0.740$ for the primary and secondary analyses. Pooling concatenates the run–criterion observations rather than averaging within-run correlations. The scan's rank routine uses ordinal ranks from a double `argsort`; equal values therefore receive distinct ranks rather than average ranks.

Table 10 lists the six runs with their displacement at both measured steps. Neither analysis matches KL across angles.

The primary analysis provides the fixed-step comparison; the secondary association is reported with its angle-ordered displacement values. Together these measurements examine how a geometric intervention relates to criterion responses across a controlled family of training signals.

F External reward-derived subspace experiment

F.1 An external reward model supplies the driving signal

We next derive a driving direction from ArmoRM (Wang et al., 2024a). This direction is fitted in the policy model's representation space; it is not an ArmoRM native weight vector. The external pipeline regresses the reward model's helpfulness, correctness, and coherence attribute scores on

Table 12: Second external reward model, matched-KL two-arm comparison. Retention is the above-reference agreement of the treated arm as a fraction of the control arm, using the fixed 0.5 reference for this secondary summary. The driving range combines the two arms' bootstrap endpoints; the scorecard range is conditional on the control point estimate. Agreement is length-adjusted; intervals use question-clustered bootstrap resampling.

readout	α^{ctrl}	α^{treat}	retention
InternLM2-derived (driving)	0.5682	0.5248	36.4% $[15.0, 67.3]$
eleven criteria, mean	0.5461	0.4953	-10.1% $[-21.9,\ \ 1.9]$

within-question-centred features, using centred targets. The normalised mean of the three fitted coefficient vectors defines u_{RM}. Training with this signal supplies the control arm.

Across the eleven evaluation readouts, geometric overlap with this direction tracks length-adjusted agreement, with reported pooled Spearman $\rho^{\perp} = +0.929$ ($n = 22$, eleven criteria and two control seeds). The reported projected overlap of `harmlessness` is 0.047.

F.2 Subspace removal evaluated on eleven criteria

Let $W_{\mathcal{H}}$ collect the external pipeline's own attribute probes fitted from HelpSteer2. Let $U_{\mathcal{H}}$ denote the basis used by that external fitting pipeline and $P_{\mathcal{H}} = U_{\mathcal{H}} U_{\mathcal{H}}^{\top}$. The matrix $Q_{\mathcal{H}}$ contains the orthonormal columns returned by the default QR factorisation of $P_{\mathcal{H}} W_{\mathcal{H}}^{\top}$. For a nonzero residual, the external intervention is given by

$$u_{\mathrm{RM}\setminus\mathcal{H}} = \frac{u_{\mathrm{RM}} - Q_{\mathcal{H}} Q_{\mathcal{H}}^{\top} u_{\mathrm{RM}}}{\| u_{\mathrm{RM}} - Q_{\mathcal{H}} Q_{\mathcal{H}}^{\top} u_{\mathrm{RM}} \|}. \tag{19}$$

The implementation uses no explicit QR rank threshold. The columns represented by $Q_{\mathcal{H}}$ are distinct from the eleven FIT2 criterion probes used for evaluation. Consequently, changes across that scorecard test transfer of the subspace intervention, rather than orthogonality imposed on its eleven readouts.

Figure 5 contrasts the two external reward models and Table 12 gives the second model's driving-readout and scorecard summaries. The reported overlaps use directions projected and normalised in the evaluation basis S. Under this convention, the cosine between the control and treated directions is 0.837, and the treated overlaps with the eleven criterion readouts range from 0.014 to 0.111. These are measured geometric properties, not constraints imposed on those readouts by the construction. The evaluation compares two control seeds (both step 2200, KL 0.4109 and 0.4529) with one treated seed (step 6000, KL 0.4248). Table 11 reports the resulting length-adjusted agreement.

For this secondary fixed-0.5 summary, agreement with the reward-model readout remains above that reference in the treated arm (0.5990, 95% CI $[0.5867, 0.6117]$, excluding 0.5), retaining 74.1% of its above-reference signal. The scorecard retains 39.3%: the gap is positive for *all eleven* criteria with every interval excluding zero, and the equally weighted aggregate is $+0.0520$, 95% CI $[+0.0467, +0.0574]$. The intervention produces lower agreement across the scorecard while preserving most of the external driving readout's above-reference signal.

The smallest measured agreement difference occurs for `harmlessness`, the criterion whose overlap increases from 0.047 to 0.092. Its difference of $+0.0129$ is approximately one quarter of the eleven-criterion mean ($+0.052$), providing a further correspondence between geometry and the response pattern.

G The chance level of the agreement statistic

The agreement credit $A(d, z)$ of Appendix D gives half credit when the policy difference ties ($d = 0$) and full credit when the two signs match, but *no* credit when the criterion readout ties while the policy difference does not. Readout ties are not negligible here. Of the $61{,}159$ same-question response pairs in the evaluation partition, 6.47% join two responses whose hidden states are identical — duplicate responses in the source data — so $z_X = 0$ for *every* criterion on those

Table 13: Retention and removal under the two references. The 0.5 column is what a fixed half-way reference would give; the 0.468 column is the empirical chance reference used for the main-table ratios.

pair	$A \to B$	R_A at 0.5	R_A at 0.468	E_B at 0.5	E_B at 0.468
1	completeness → insightfulness	82.1%	84.0%	37.1%	32.7%
2	logical efficiency → conciseness	99.0%	99.2%	99.0%	**54.2%**
3	logical robustness → comprehension	75.6%	79.2%	39.8%	34.1%
4	logical correctness → commonsense	87.6%	89.8%	35.9%	30.2%
5	readability → factuality	96.4%	96.9%	30.8%	24.5%

Table 14: The aggregate under five analysis conditions. Every row is positive in all included criterion pairs, with every interval excluding zero. The exposed-question layer gives $T = +0.0636$ $[+0.0595, +0.0677]$, differing from the unexposed layer by 0.0014.

condition	what it varies	questions	T	95% CI
main analysis	—	583	+0.0633	[+0.0596, +0.0670]
length-adjusted	criterion scores residualised on length	583	+0.0544	[+0.0508, +0.0582]
unexposed questions	questions outside the DPO candidate pool	112	+0.0622	[+0.0537, +0.0708]
Qwen2.5-0.5B, four pairs	model scale, refitted probes and basis	583	+0.0633	[+0.0589, +0.0676]
Qwen2.5-1.5B, same four	pair subset, for like-for-like comparison	583	+0.0561	[+0.0518, +0.0604]

pairs. Under a reference model with independent, equiprobable, non-tied policy-change signs, the expected agreement is approximately

$$(1 - 0.0647) \times \tfrac{1}{2} + 0.0647 \times 0 \;=\; 0.46765 \;\simeq\; 0.468, \tag{20}$$

and not $\frac{1}{2}$. If policy ties occur, the null expectation instead includes their half credit: $\frac{1}{2}\Pr(d = 0) + \frac{1}{2}\Pr(d \neq 0, z_X \neq 0)$, assuming equiprobable policy signs conditional on non-ties. For the non-tied random-sign reference, the value is 0.4677–0.4681 across the eleven criteria, effectively a constant, because the ties are driven by duplicate responses rather than by any criterion's own resolution.

Reference-invariant contrasts and reference-specific ratios. The aggregate T and the between-arm agreement contrasts are invariant to the reference: $T = +0.0633$ and their confidence intervals are unchanged. Retention ratios and their intervals depend on the chosen reference. The reference enters only the ratios R_A and E_B, which express agreement above chance as a fraction of the control arm's agreement above chance. Table 13 gives both versions. Retention R_A is almost unchanged, since α_A sits far above either reference. Removal E_B falls by about five points in four pairs and sharply in pair 2, whose treated arm happens to sit at 0.5004: measured against 0.5 this looks like near-total removal, while against the chance level of 0.468 the treated arm remains 0.032 above chance and the removal is 54.2%. The main-table ratios use the chance-level version. External-signal fixed-0.5 ratios, the census confirmation's registered comparison, and the checker-selection endpoint retain their explicitly stated references.

Why the estimand is not changed. Excluding all criterion-tied pairs would make the non-tied random-sign reference exactly $\frac{1}{2}$, but it would also define a different α and hence a different T from the one fixed before training. We therefore leave the statistic as pre-registered and correct only the reference used to interpret ratios.

H Robustness diagnostics

Table 14 gives the five conditions and their aggregates.

Length-adjusted agreement. The control arm's preference changes agree with response-length rankings at rates of 0.679 to 0.750 across pairs. The between-arm difference in length agreement

is at most 0.0295. For criterion pair-score differences z_X and response character-count differences L, both over the same $61{,}159$ EVAL pairs, the diagnostic uses

$$z_X^{\perp} = z_X - \frac{L^\top z_X}{L^\top L} L. \tag{21}$$

For $L^\top L > 0$, this is a through-origin least-squares residual computed on the evaluation pairs themselves. Replacing z_X by $z_X^{\perp}$ in the agreement computation leaves $T^{\perp} = +0.0544$, 95% CI $[+0.0508, +0.0582]$, positive in all five pairs with every interval excluding zero. Here $\perp$ denotes this length adjustment, distinct from the coefficient-space intervention. The angle and external-signal analyses also report length-adjusted agreement.

Agreement on unexposed questions. The DPO candidate pool is drawn from 80% of all questions, chosen independently of our partitioning, so 471 of the 583 EVAL questions appear as training candidates. Both arms draw from the same candidates with an arm-independent seed, so candidate exposure is shared across arms. We assess the effect separately on exposed and unexposed questions: all 583 questions give $T = +0.0633$ $[+0.0596, +0.0670]$; the 112 questions with no exposure of any kind give $T = +0.0622$ $[+0.0537, +0.0708]$, positive in all five pairs with every interval excluding zero; the 471 exposed questions give $T = +0.0636$ $[+0.0595, +0.0677]$. The clean and exposed layers differ by 0.0014. The clean subset supports the main result on questions outside the DPO candidate pool.

Replication at a second model scale. Reusing the same five pairs and orientations on Qwen2.5-0.5B, which refits its own probes and measurement basis, four pairs admit a matched-KL comparison. On exactly those four pairs the aggregate is $T = +0.0633$ $[+0.0589, +0.0676]$ at the smaller scale and $T = +0.0561$ $[+0.0518, +0.0604]$ at the larger, positive in every pair at both scales.

Implementation checks. Every adapter is checked to have loaded by comparing its log-probabilities against the base model (16/16 pass). The agreement statistic is unbiased to within 6×10^{-4} on synthetic data with known correlation.

Observed run-to-run variation. Repeating a training run with an identical seed and an identical data order produces relative differences in KL of 2.5% to 17.2% across checkpoints. These observations quantify variation in this repeat; they do not establish a universal noise floor or identify its source. The relative matching tolerance (≤ 0.25) exceeds this observed range. The accepted between-arm gaps (≤ 0.10 absolute) remain measured differences in KL; satisfying these thresholds does not establish equal displacement.

I Independent-seed replication of the trajectory prediction

What was frozen, and when. The protocol was frozen on 2026-09-22, before any arm of seed 20260922 was trained and before any endpoint of that seed was read. The model, the two pair directions, the training and evaluation code, and the frozen EVAL inputs are each hash-locked by SHA-256. It also records what was known at freeze time: on the development seed, pair 4's contrast was -0.0576 at step 100 and $+0.0171$ at step 400, and pair 1 was positive at both. Those values motivate the prediction and are not part of the decision.

Prediction and decision rule. Four arms were trained — control and treated for pairs 1 and 4 — to a fixed 3000-step cap with checkpoints at $100, 200, 400, 700, 1100, 1600, 2200, 3000$, under ordinary DPO with the main experiment's hyper-parameters. No result-dependent extension, stopping, arm substitution or repeated seed was permitted. The endpoint is the original EVAL population with the original frozen linear coefficients and tie handling; intervals are pointwise, from 2000 question-clustered resamples. The prediction is the conjunction $\Delta_B(4, 100) < 0$, $\Delta_B(4, 400) > 0$, $\Delta_B(1, 100) > 0$, $\Delta_B(1, 400) > 0$, and it is declared supported only if every one of the four intervals lies entirely on its predicted side. A separate dose qualifier permits the phrase *near-matched dose* at pair 4 step 400 only if the two arms' KL ratio lies in $[0.8, 1.25]$.

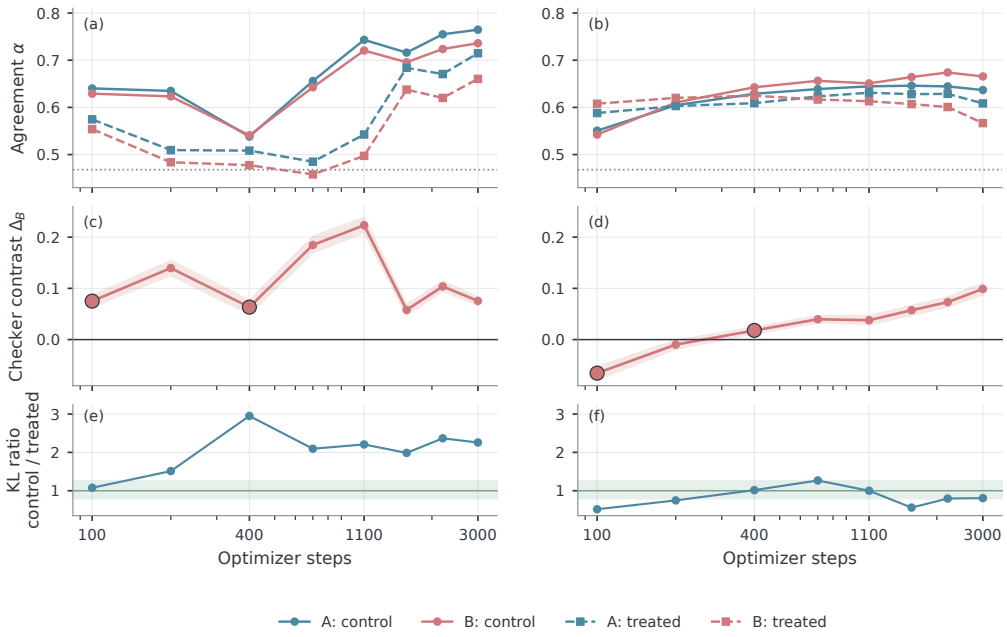

Figure 8: **A frozen sign prediction reproduces on an independently trained seed. Top:** agreement of both criteria in both arms across training. The gap that Δ_B measures is the vertical distance between the two checker curves; the driving criterion's two curves stay closer together. Dashed horizontal line, empirical chance. **Bottom:** the contrast itself, with pointwise question-bootstrap intervals. Filled circles are the two steps whose signs were fixed before seed 20260922 was trained; open triangles, offset left for visibility, are the same two steps on the development seed that motivated the prediction. Pair 4 crosses zero between steps 100 and 400 on both seeds. Shaded columns mark the four of sixteen checkpoints at which the two arms' conditional KL agree to within $[0.8, 1.25]$ — including pair 4's crossing, at a ratio of 1.016. The unshaded steps are descriptive: at equal step counts the arms are not in general at equal displacement.

Outcome. All four intervals fall on the predicted side, and the dose qualifier is met at the crossover. Figure 8 plots both trajectories; the full eight-point table follows. The two pre-declared steps are marked and the remaining six are descriptive. Pointwise intervals are not a simultaneous band.

Reading the trajectory. The pattern, including pair 4's change of sign, reproduces on a seed whose endpoints were unseen when the prediction was fixed. That is a statement about temporal reproducibility: nothing here identifies which component of the update produces the crossover. The KL column constrains how the remaining checkpoints may be read — pair 1's arms differ by a factor of two to three at most steps — so these fixed-step contrasts stand apart from the matched-KL estimates of Table 2 and none of them enters T. One additional training seed, conditional on a fixed dataset and evaluator, is not an estimate of seed-to-seed variability.

J READOUT AND OPTIMISATION SCOPE

Linear readouts make the construction and its scoring sensitivity explicit, while DPO provides a controlled test of the resulting training directions. The reported matched-KL evidence concerns this setting. Pre-specified tests with other preference objectives produced no admissible KL-matched comparison, so they supply no estimate of detection separation. Matching is defined by policy displacement, not by equal optimiser-step budgets.

The nonlinear extensions used the same targets and a held-out SELECT validity gate fixed before evaluation. The required criterion pairs did not pass that gate, so those extensions do not establish

Table 15: Independent-seed trajectory, all eight checkpoints. Starred steps are the two whose signs were fixed before training.

pair	step	Δ_B	pointwise 95% CI	KL ctrl / treat	ratio
1	100*	+0.0751	[+0.0643, +0.0858]	0.0033 / 0.0031	1.08
1	200	+0.1396	[+0.1258, +0.1539]	0.0067 / 0.0045	1.51
1	400*	+0.0634	[+0.0501, +0.0761]	0.0341 / 0.0115	2.95
1	700	+0.1847	[+0.1699, +0.2010]	0.0967 / 0.0461	2.10
1	1100	+0.2232	[+0.2090, +0.2379]	0.1530 / 0.0693	2.21
1	1600	+0.0580	[+0.0479, +0.0679]	0.1764 / 0.0888	1.99
1	2200	+0.1037	[+0.0947, +0.1136]	0.2896 / 0.1223	2.37
1	3000	+0.0755	[+0.0679, +0.0827]	0.3761 / 0.1665	2.26
4	100*	−0.0655	[−0.0776, −0.0532]	0.0028 / 0.0054	0.52
4	200	−0.0099	[−0.0186, −0.0015]	0.0082 / 0.0110	0.75
4	400*	+0.0178	[+0.0117, +0.0238]	0.0570 / 0.0561	1.02
4	700	+0.0397	[+0.0333, +0.0463]	0.0815 / 0.0643	1.27
4	1100	+0.0378	[+0.0297, +0.0463]	0.1819 / 0.1819	1.00
4	1600	+0.0573	[+0.0477, +0.0662]	0.2480 / 0.4425	0.56
4	2200	+0.0732	[+0.0641, +0.0827]	0.3183 / 0.3998	0.80
4	3000	+0.0989	[+0.0881, +0.1098]	0.4296 / 0.5322	0.81

nonlinear-readout generality. This is a boundary on the evidence, not evidence that a valid nonlinear checker would share the measured separation.

K A MIS-SPECIFIED CHECKER READOUT

Question and pre-registration. The construction is defined using the checker criterion's readout. How accurate must that readout be? The recipe for producing a mis-specified readout, the guardrail, the pair selection and the decision table were frozen before any mis-specified direction was produced and before any arm was trained.

Recipe. For each pair the checker readout is refit on the FIT2 partition with three changes and no others: a 50% subsample without replacement, ridge coefficient 10.0 in place of 1.0, and fitting seed 20260918. The measurement basis, standardisation and sign-alignment convention are unchanged. Let b and b' denote the normalised projections of the original and refitted checker coefficients, and let $c' = |a^\top b'|$. The blind direction is then built from b' rather than b. A guardrail excludes any pair with $\cos(b, b') > 0.98$, on the grounds that such a readout is not in fact mis-specified; no pair triggers it. Measurement always uses the true readout — using b' to measure would be circular and is prohibited by the protocol.

Pairs are the top three of the main table by Δ_B, fixed before any result of this round existed. Control arms are reused from the main experiment: u_A does not depend on b, so the mis-specification enters only the treated direction and the control arm is literally the same training run.

pair	$A \to B$	$\cos(b, b')$	c	c'
1	*completeness → insightfulness*	0.876	0.560	0.604
3	*logical robustness → comprehension*	0.890	0.584	0.515
4	*logical correctness → commonsense understanding*	0.829	0.621	0.521

Result. All three pairs match at the primary KL point within the usual tolerances, with between-arm gaps of 0.060, 0.072 and 0.044, and all three clear the retention gate (0.904, 0.907, 0.989). Figure 9 shows the outcome.

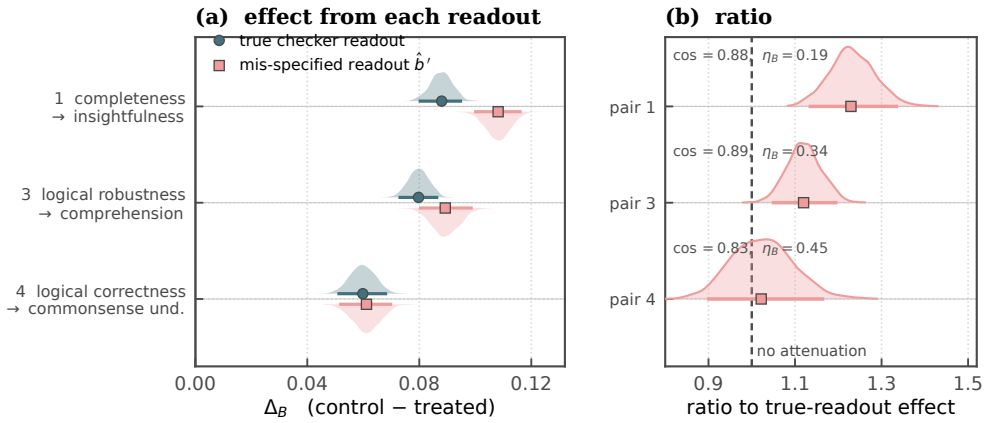

Figure 9: **An approximate checker readout is enough.** **(a)** The bypassed-criterion contrast at matched KL for each pair, with the direction built from the true checker readout and from one deliberately refit on half the sample at a ten-fold ridge coefficient. Shaded shapes are the 2000 question-clustered bootstrap distributions, bars the 95% intervals. **(b)** Their ratio, recomputed inside each shared resample so that numerator and denominator move together; the dashed line marks equality, and each row carries the cosine between the two readouts and the residual sensitivity η_B of the mis-specified direction, both computable before training. Two ratio intervals lie entirely above equality and the third straddles it. We read this as survival, not as improvement.

pair	Δ_B (mis-specified)	Δ_B (true readout)	ratio
1	$+0.1082$ $[+0.0996, +0.1166]$	$+0.0880$	1.229 $[1.131, 1.338]$
3	$+0.0893$ $[+0.0799, +0.0991]$	$+0.0797$	1.120 $[1.046, 1.199]$
4	$+0.0611$ $[+0.0514, +0.0703]$	$+0.0598$	1.022 $[0.896, 1.168]$

The load-bearing criterion — a positive Δ_B with an interval excluding zero on at least two pairs — is met on all three. Ratios are computed within each of the 2000 shared resamples, recomputing numerator and denominator in the same isolated run so that both sides use one bootstrap and the seed-0 cell on each side; the median ratio is 1.120.

Reading. An approximate checker readout is enough. All three mis-specified directions retain a positive contrast: two ratio intervals lie entirely above 1 and the third covers it. Whether mis-specification *helps* is not something three pairs, one seed and no mechanism can settle; the supported statement is that all three contrasts remain positive. Nor does this speak to a black-box setting: b' is a perturbation of a white-box readout, not the absence of one. This round is reported separately and is not pooled into T.

L A THIRD MODEL SCALE: QWEN2.5-7B

What was fixed in advance. The protocol, the gates and the decision table for this round were frozen before any readout was fit on 7B representations and before any pair was selected. A dated amendment adopted the main experiment's pair-selection rule rather than the census rule before any Δ_B was computed. The 7B experiment also differs in batch size and checkpoint selection, as specified below.

Setup. Qwen2.5-7B base (28 layers, $d = 3584$), readouts fit at layer 17 on the same FIT2 partition, with the same ridge coefficient and the same 100-dimensional measurement basis, refit on this model's own hidden states over the same FIT2 partition. Of the 110 ordered pairs, 41 are admissible; the frozen rule selects *completeness* → *logical efficiency*, with in-span alignment $c = 0.5687$ and $\delta = 34.66°$. One pair is trained, for compute reasons; the rule, not the outcome, determines which.

Batch-size difference. Both arms use batch size 8 rather than 16. At 7B the policy and reference weights alone occupy 30.4 GB, which fills a 32 GB card irrespective of batch size, so the pre-registered ladder $16 \to 8 \to 4$ is descended one step. Both arms use the same value. Learning rate, β, LoRA rank, the KL target and every decision criterion are unchanged.

Matching. The implemented selector minimises the sum of the two checkpoint step counts among pairs satisfying the individual KL tolerances and arm-gap bound. This differs from the main experiment's nearest-target rule. The control arm qualifies at steps 700, 2200 and 3000; the selected pair uses control step 700 ($\mathrm{KL} = 0.408$). The treated arm does not reach 0.45 within the original schedule ($\max \mathrm{KL} = 0.241$) and is extended once on a KL-only basis, as the protocol permits; within the extension it qualifies at steps 2200, 7600 and 13000, and the joint selector uses 2200 ($\mathrm{KL} = 0.345$). The between-arm gap is 0.063, inside the 0.10 ceiling, and both selected checkpoints clear the degeneration gate (perplexity ratios 1.41 and 1.43 against a threshold of 10).

Both KL trajectories are non-monotone. The control arm reads 0.010, 0.141, 0.190, 0.408, 0.245, 0.252, 0.557, 0.498 over the schedule; the treated arm reads 0.002, 0.018, 0.069, 0.140, 0.127, 0.154, 0.345, 0.291 and then 0.329, 1.135, 0.525, 0.627, 0.345 over the extension. Because training here is not bitwise reproducible (Appendix H), the extension is treated as one complete trajectory and is never spliced with the original run.

Result. Measurement uses the main protocol's raw agreement endpoint: EVAL partition, same-question pairing, and question-clustered bootstrap with $B = 2000$ and seed 20260911 (8741 responses, 583 questions, $61{,}159$ same-question pairs), and the empirical chance level 0.468.

	control arm	treated arm
α_A (driver, *completeness*)	0.7586 $[0.7479, 0.7690]$	0.7398 $[0.7296, 0.7499]$
α_B (checker, *logical efficiency*)	0.5688 $[0.5535, 0.5830]$	0.5369 $[0.5229, 0.5505]$

The registered power gate is met: both control-arm intervals lie above 0.5. The load-bearing criterion is met: $\Delta_B = +0.0319$ with 95% CI $[+0.0254, +0.0379]$, excluding zero. The retention gate is met with a ratio of 0.975. In the normalisation used throughout, the driver criterion keeps 93.6% of its above-chance agreement, inside the 79.2%–99.2% range observed at 1.5B, while 31.6% of the checker criterion's above-chance agreement is removed.

Scope of this round. This pair is not one of the five at 1.5B: each model's geometry is its own, and the frozen rule selects within it. What transfers to a scale $4.7\times$ larger is therefore the construction under a criterion fixed in advance. A scale trend would need more than the two points available, and Δ_B here is not pooled into T, which is defined over the five pairs at 1.5B.

M SINGLE-ARM RESPONSE AND THE TWO-ARM ESTIMAND

A constructed direction and the policy obtained by training along it are distinct objects. The two-arm design in Section D measures the change in agreement relative to a control trained at matched KL displacement.

As a single-arm illustration, interventions targeting ArmoRM as the bypassed readout were tested in four configurations at matched KL. Three have absolute cosine overlap of 0.037 to 0.064 with the reward readout and post-training agreement of 0.514 to 0.556; the fourth, built from an earlier and less complete construction, retains overlap 0.424 and reaches 0.611. Every interval excludes 0.5. Small directional overlap can therefore coexist with above-reference agreement after training, while the ordering across configurations still follows overlap. This motivates estimating the between-arm difference under shared candidate data and matched displacement. For identical policy log-probability arrays, the two-arm difference is zero by the definition of the agreement statistic.

Table 16: Census of the construction over all 110 ordered criterion pairs on the base model. The construction is defined for every pair. η_A is the driving criterion's retained scoring sensitivity, η_B the bypassed criterion's. No training is involved.

	min	10%	25%	50%	75%	90%	max
c	0.008	0.062	0.234	0.396	0.530	0.596	0.797
η_B	0.003	0.039	0.088	0.139	0.202	0.435	5.073
η_A	0.563	0.756	0.808	0.894	0.963	0.999	1.001

Table 17: Where the five analysed pairs fall in the census distribution. Percentiles are computed over all 110 pairs; a higher percentile means a larger value of that quantity.

#	$A \to B$	c	pct	η_B	pct
1	completeness → insightfulness	0.560	80%	0.181	71%
2	logical efficiency → conciseness	0.359	42%	0.241	80%
3	logical robustness → comprehension	0.584	87%	0.146	55%
4	logical correctness → commonsense	0.621	95%	0.197	74%
5	readability → factuality	0.435	58%	0.027	6%

N How common is the construction? A census over all ordered pairs

`harmlessness` anchors the low-removal end of this relationship. Its predicted overlap change is small in all five pairs ($+0.0506$, -0.0194, $+0.0306$, -0.0112, -0.0063). In the three pairs where it is the smallest such change among all non-driving criteria (pairs 3, 4 and 5), its measured agreement difference ranks first, second and first of ten ($+0.0131$, $+0.0143$, -0.0007). In pairs 1 and 2 its predicted change is larger and its rank is correspondingly lower. The correspondence is therefore between the predicted and the measured ordering, not an unconditional minimum.

The five criterion pairs analysed in the main text were chosen by the selection rules fixed before training (Appendix D). This appendix asks how those five sit inside the full set of possibilities. We recompute the construction geometry for every ordered pair of the eleven FLASK criteria, $11 \times 10 = 110$ pairs in total. Everything here is evaluated on the base model with the frozen probe coefficients and measurement basis; no training is involved, and the census itself uses no agreement statistic.

What is computed. For an ordered pair (A, B) we form a, b, s, c and $u_{A\setminus B}$ exactly as in Appendix C, and report the residual scoring sensitivity

$$\eta_B = \frac{\left|u_{A\setminus B}^{\top}(W_B/\sigma)\right|}{\left|a^{\top}(W_B/\sigma)\right|}, \tag{22}$$

the treated direction's sensitivity on criterion B relative to the control direction's, in the raw-state coordinates where scoring actually happens. The ratio requires a nonzero denominator; a zero baseline sensitivity does not define a retention ratio. η_A is the same quantity for the driving criterion. As a check, the recomputation reproduces the frozen c and δ of the five analysed pairs to machine precision.

The construction is available broadly. It is defined for all 110 pairs (Table 16). The driving criterion retains more sensitivity than the bypassed one in 105 of them. The median η_B is 0.139: for a typical pair, removing the overlap leaves the second criterion with about one seventh of the scoring sensitivity it had under the unmodified direction. Counting by threshold, $\eta_B \le 0.25$ holds for 89 pairs, $\eta_B \le 0.10$ for 35, and $\eta_B \le 0.05$ for 14.

The analysed pairs are not the most favourable ones. The selected pairs tend to have large construction overlap in this census (Table 17): the median percentile of c across the five is 80%. Residual sensitivity runs the other way. The median η_B percentile of the five is 71%, so on these pairs the

Table 18: Confirmation round. Selection by ascending η_B; two-arm DPO and measurement identical to the main experiment. Pair 7's arms required one schedule extension under the KL-only rule; pair 6's did not. R_A and E_B use the fixed 0.5 reference of this round's registered comparison; Table 2 uses the empirical chance reference.

#	$A \to B$	η_B	c	$\alpha_B^{\text{ctrl}} \to \alpha_B^{\text{treat}}$	Δ_B	95% CI	R_A / E_B
6	`conciseness` → `comprehension`	0.0031	0.361	.6038→.6117	$-$**.0079**	$[-.0188, +.0021]$	81.0% / −7.6%
7	`completeness` → `logical robustness`	0.0140	0.518	.6374→.5804	$+.0569$	$[+.0493, +.0644]$	93.2% / 41.5%

construction removes *less* of the bypassed criterion's scoring sensitivity than it does on a typical pair, and the blind-spot effects of Section 4.3 are obtained under that weaker removal. Of the ten pairs with the smallest η_B, one belongs to the main five; two others were trained in the confirmation round below.

A census of directions, not of effects. The quantities here characterise the construction on the base model across all pairs. Agreement is never measured, and η_B is a property of directions rather than of a trained policy, so the census alone establishes no training effect; the additional two-arm tests are reported below. The census is reported so that the position of the analysed pairs within the full distribution is visible rather than left implicit.

N.1 Separating construction sensitivity from training response

Two additional pairs test whether lower residual scoring sensitivity η_B predicts stronger empirical detection separation. Selection and the prediction were fixed before their two-arm training comparisons.

Selection and prediction, both frozen before training. Pairs were ranked by ascending η_B over the 110 ordered pairs, excluding `harmlessness`, excluding the five analysed pairs, and requiring the orientation and non-redundancy conditions of Appendix D; the first two were taken. This is a different rule from the one that produced the five. The prediction was that a construction an order of magnitude more complete would remove at least as much of the bypassed signal as the five do, whose median E_B under this round's fixed 0.5 reference is 37.1%. We recorded in advance that the trend already visible across the five pointed the other way.

The prediction is refuted. Table 18 gives both cells. Pair 6 has the more complete construction by a factor of four and the smallest η_B of any pair we have trained, and its interval covers zero: the predicted positive effect is not supported. Pair 7, with four times the residual sensitivity, shows a clear one. The ordering is the reverse of the prediction, and it agrees with the five-pair trend $\rho_{\text{Spearman}}(\eta_B, E_B) = +0.6$. Two rounds therefore point the same way: η_B **does not order the effect, and a more complete construction is not a stronger one**. We report η_B as a description of the construction and make no predictive claim for it in the main text.

What pair 7 does add. Pair 7 was selected by a different ranking rule while retaining the orientation and eligibility constraints, and the effect appears at $R_A = 93.9\%$ and $E_B = 33.6\%$ using the empirical 0.468 reference, both inside the five-pair range (79.2%–99.2% and 24.5%–54.2%). The effect is therefore not an artefact of the original selection rule. This is what the round establishes; it leaves the registered prediction unsupported; both selected pairs appear in Table 18.

Pair-level interpretation. Both selected pairs are reported individually to preserve their distinct responses. Their selection rule differs from the five-pair experiment, so their results are separate from T. The two pairs reuse dimensions that appear among the five (`conciseness`, `comprehension`, `completeness`, `logical robustness`), so they are *not* independent of them, and the once-per-dimension condition holds within the original five only. Finally, pair 7's KL trajectory is markedly non-monotone (the treated arm passes 0.2016, 0.2960, 0.5005, 0.3833, 0.5041), so its matched checkpoint is selected from a noisier trace than the main experiment's; the run-to-run floor of Appendix H applies.

Table 19: Single-checker removals with the smallest residual scoring sensitivity for each driving criterion. η_A is the fraction of the driving criterion's sensitivity the evading direction retains. Base model only.

driving criterion A	targeted checker	η	η_A
conciseness	comprehension	0.003	0.904
readability	completeness	0.003	0.875
completeness	logical robustness	0.014	0.826
commonsense understanding	factuality	0.019	0.760
insightfulness	logical robustness	0.023	0.944
comprehension	completeness	0.033	0.563
factuality	completeness	0.037	0.776
harmlessness	logical correctness	0.057	0.994
logical efficiency	logical correctness	0.065	0.805
logical robustness	insightfulness	0.075	0.949
logical correctness	logical efficiency	0.095	0.816

Table 20: Adding checkers. Medians over the eleven driving criteria, with checkers added greedily. The maximum residual sensitivity over eligible non-driving criteria falls; median driving sensitivity decreases from 0.826 to 0.783.

$\lvert C\rvert$	$\max_{X\in C}\eta_X$	$\max_{X\in\mathcal{E}_A}\eta_X$	η_A
1	0.033	0.916	0.826
2	0.039	0.793	0.796
3	0.034	0.658	0.794
4	0.051	0.625	0.783

O CHOOSING CHECKERS: A COVERAGE AUDIT

The construction of Section 3 can be read in the other direction. Given a criterion A that drives training and a set C of criteria used to check the result, the direction that evades all of C while retaining as much of A as possible is

$$u_C = \frac{(I - P_C)\, a}{\|(I - P_C)\, a\|}, \qquad \|(I - P_C)\, a\| > 0, \tag{23}$$

where P_C projects onto the span of the checkers' normalised projected coefficients. This appendix measures the residual scoring sensitivity of checker sets against directions constructed to evade them. Everything is computed on the base model with the frozen coefficients; no training is involved and no agreement statistic appears.

What is reported. Coefficient orthogonality is not invisibility: scoring happens in raw-state coordinates, so the quantity that matters is the residual scoring sensitivity $\eta_X(C) = |u_C^\top (W_X/\sigma)|/|a^\top (W_X/\sigma)|$ of Appendix N. Let $\mathcal{E}_A$ be the non-driving criteria passing the denominator guard below. We report $\max_{X\in C}\eta_X$ (residual sensitivity of the chosen checkers), $\max_{X\in\mathcal{E}_A}\eta_X$ (the maximum over all eligible non-driving criteria, including chosen checkers) and η_A (how much of the driving criterion survives). Checkers are added greedily to minimise $\max_{X\in C}\eta_X$. As a check, the $|C| = 1$ values reproduce the census of Appendix N to machine precision.

Geometric sensitivity and measured detection. For a fixed driving criterion the largest and smallest residual sensitivities differ by a factor of about thirty (Table 19 lists the minima; the full 110-entry matrix is the census of Appendix N). The quantity is available before training. It does not, however, translate into stronger retained detection when a checker is actually chosen by it and the result is measured: Section 5 tests exactly that and finds the geometric rule below the agreement-based baseline, for a reason identified there. The statements of this appendix are about the geometry of the base model and should be read as such.

Table 21: Reversed orientation, KL trajectories. The treated arm has lower KL at most checkpoints, but higher KL at step 400 in both seeds. No pair–seed cell satisfies all matching rules, so no agreement statistic is computed from these runs.

	arm	100	200	400	700	1100	1600	2200	3000
seed 0	control	0.0021	0.0053	0.0287	0.1589	0.2190	0.3018	0.3892	0.3196
	treated	0.0015	0.0052	0.0383	0.0849	0.1056	0.1621	0.3167	0.2790
seed 1	control	0.0023	0.0057	0.0297	0.1594	0.2241	0.2819	0.4801	0.2936
	treated	0.0016	0.0057	0.0395	0.0836	0.1050	0.1597	0.3069	0.2487

Adding checkers does not remove the direction. Going from one checker to four leaves the driving criterion's retained sensitivity at $\eta_A = 0.783$, compared with 0.826 for one checker (Table 20). What does change is the footprint elsewhere: the largest residual over eligible non-driving criteria falls from 0.916 to 0.625. Defending against more criteria therefore makes the evading direction quieter across the scorecard rather than more conspicuous, while leaving most of the driving signal intact. This agrees with Section 4.5, where removing the span of all eleven criteria retains external driving-readout agreement of 0.5248 after removing the eleven-readout span in the InternLM2-derived experiment.

A ratio guard. η is a ratio, and its denominator $|a^\top (W_X/\sigma)|$ is near zero for criteria the control direction barely moves. With `readability` driving, `harmlessness` has denominator 1.5×10^{-4} and numerator 7.8×10^{-4}: both are small, and their ratio of 5.07 reflects a small denominator. Criteria whose denominator falls below 10% of the median for that driving criterion are excluded from the $\max_{X\notin C}$ statistic. Ten pairs are excluded, and all ten involve `harmlessness` as the checked criterion; the ten are enumerated in the released measurements.

A pre-training ordering. This section orders checkers by a quantity computable before training, and bounds what adding checkers achieves under this construction. Whether a low η implies detection after training is a separate question: η is a property of directions on the base model, and no training result enters this section.

P Reversed orientation: a control that yielded no usable cell

The orientation rule assigns A to the side with higher held-out validity, and we make no claim that the construction is symmetric under exchanging the two criteria. To probe that asymmetry we ran the exchanged version of pair 3, taking `comprehension` as the driving criterion and `logical robustness` as the bypassed one. The removable component c is symmetric under exchange, so the reversed pair carries the same $c = 0.5837$ and the same $\delta = 35.71°$ as the forward pair.

The comparison was never formed. Under the same matching rule applied everywhere in this paper, neither seed produced a usable cell: the treated arm reached a maximum KL of 0.317 (seed 0) and 0.307 (seed 1) against a target of 0.45, absolute deviations of 0.133 and 0.143 — relative deviations of 0.296 and 0.318 — against a tolerance of 0.25 relative and 0.15 absolute. Extending the schedule once, under the KL-only extension rule, did not change this. Both seeds are therefore left blank, and no reversed estimate is reported. No effect estimate is available for the reversed orientation.

What the trajectories still show. Table 21 records a descriptive fact that does not depend on any agreement statistic. In the forward orientation the treated arm displaces the policy *faster* than the control at matched step count (at step 700, 0.301 against 0.176); in the reversed orientation it is *slower* at step 700 (0.085 against 0.159 at the same step). The rate at which training along $u_{A\setminus B}$ moves the policy therefore depends on which member of the pair drives, even though c does not. This is a single pair, observed after the forward result was known, and we report it as an observation rather than a finding.

Table 22: Aggregate over nested subsets ordered by human–model grounding. Every threshold is reported. The aggregate remains positive at every reported threshold.

threshold on $\min(\mathrm{HM})$	pairs included	T
≥ 0.20 (all five)	1,2,3,4,5	$\mathbf{+0.0633}$
≥ 0.30 / 0.40 / 0.50	1,2,3,4	$+0.0684$
≥ 0.55	1,3,4	$+0.0745$
≥ 0.60 / 0.70 / 0.80	4	$+0.0606$

Table 23: Regularisation path at $\gamma = 0$. As the penalty grows the readouts lose effective dimensionality, their subspaces overlap more, and the fraction of cells admitting a construction that preserves the driving readout falls to zero. Here d_{eff} is the effective dimensionality of the fitted readout subspace, O the overlap between the two readout subspaces, retention the within-question rank correlation between the residual-direction and original driving scores, and usable fraction the share of cells with that correlation at least 0.55. This retention is distinct from η_A.

λ	$\tilde{\lambda}$	usable fraction	median retention	d_{eff}	overlap O	first principal angle
30	0.048	0.25	0.441	1.33	0.473	22.6°
300	0.48	0.17	0.336	1.28	0.531	19.4°
3000	4.8	0.17	0.259	1.21	0.660	13.2°
30000	48	0.00	0.136	1.12	0.735	7.3°

Q Robustness across human–model agreement thresholds

Our criteria are scored by a model judge under the FLASK protocol, and the criteria differ in how well those scores track human ratings. A natural concern is that the effect might be carried by the pairs whose criteria are least human-grounded. We test this by stratifying on the human–model score correlations reported by Ye et al. (2024) for each criterion, taking $\min(\mathrm{HM})$ for a pair to be the smaller of its two criteria's correlations, and recomputing the aggregate over nested subsets.

This analysis is post hoc. It was introduced after Δ_B was unblinded, and the external metadata it uses was obtained after the results were known. It does not enter the main table, and none of the subset values below is a headline number.

Table 22 supports robustness to the human-grounding thresholds examined: removing the least human-aligned criteria preserves the aggregate effect. This agrees with the pre-registered leave-one-pair-out analysis, where dropping pair 5 gives $+0.0684$ and dropping pair 2 gives $+0.0673$.

These post-hoc subsets describe robustness across thresholds; they do not establish a monotone relationship between human grounding and effect size: pair 4 has the highest $\min(\mathrm{HM})$ at 0.816 and the third-largest Δ_B. The primary estimate remains $T = +0.0633$ over all five pairs.

R When does the construction exist? A regularisation path

The construction requires two readouts that are similar enough to share a removable component, but not so similar that removing it destroys the driving readout as well. That condition is a property of readout geometry, not of the criteria's names, which raises the question of what moves a pair of readouts into or out of it. This appendix reports a pre-registered path experiment answering that question for one lever: the ridge penalty used to fit the readouts.

Probe family and fitting path. The cells here are ArmoRM's attribute heads grouped by source, twelve usable cells, all refitted with identical code on the same hidden states, the same prompt-level split and the same preprocessing; only the ridge penalty λ and a whitening exponent γ vary. In measurement-basis coordinates, let X_S be the within-question-centred features and let $C_f = X_{S,f}^\top X_{S,f}/n_f = Q_f \Lambda_f Q_f^\top$ be their fitting-set covariance eigendecomposition. The transformed rows

are $X_\gamma = X_S Q_f \Lambda_f^{-\gamma/2} Q_f^\top$. The FLASK readouts of the main experiment are fitted on different data and are not part of the path. **This is therefore a different probe family on a different source, and it is exploratory**: it constrains the mechanism, not the five-pair estimate. We report $\tilde{\lambda} = \lambda/(n_f \cdot \overline{\mathrm{var}})$, where $\overline{\mathrm{var}} = \mathrm{tr}(X_{\gamma,f}^\top X_{\gamma,f})/(n_f m)$ and m is the measurement-subspace dimension, alongside λ because the same numerical penalty is not the same strength across models.

The chain holds in every link. The pre-registered hypothesis was that a stronger penalty lowers effective dimensionality, which raises the overlap between the two readout subspaces, which leaves less of the driving readout after removal. Over the 21 (γ, λ) points the six rank correlations, computed over the 21 path-level summaries (four of which appear in Table 23), are $\rho(\log\lambda, d_{\mathrm{eff}}) = -0.826$, $\rho(\log\lambda, O) = +0.722$, $\rho(d_{\mathrm{eff}}, O) = -0.899$, $\rho(O, \text{retention}) = -0.927$, $\rho(O, \text{usable fraction}) = -0.753$, and, as the pre-registered control, the weaker $\rho(\text{held-out } \rho, \text{usable fraction}) = +0.661$. All six match the predicted sign, and the control is weaker than the mechanism variable as specified in advance. The 21 points lie on a path and are not independent, so we report no p-values.

Whitening removes the effect exactly. Varying λ from 30 to 30000 changes the usable fraction by -0.250 at $\gamma = 0$, by -0.167 at $\gamma = 0.5$, and by 0.000 at $\gamma = 1$, where every recorded quantity is constant across the whole path. Under exact full whitening, $X_{1,f}^\top X_{1,f} = n_f I$, so $(X^\top X + \alpha I)^{-1}$ is a scalar rescaling, and ridge changes the length of a coefficient vector without changing its direction, while the construction depends only on direction. The observed decrease in path dependence as γ increases is consistent with this covariance-anisotropy mechanism. In this batch the covariance restricted to the measurement subspace has condition number 3.3×10^2.

Removed-variance prediction. We had pre-registered the prediction that the removed component lies along low-variance directions, and that its relative variance would correlate negatively with the construction's efficiency. The observed correlations are $+0.524$ and $+0.554$ after partialling out removed energy: the sign is the opposite of the prediction, and we record the hypothesis as refuted. Ten of the twelve usable cells involve the same readout family, so we do not develop the reversed reading either. The reported correlations use the unnormalised reduction and partial out removed energy, separating the variance association from denominator coupling.

What the boundary accounts for. The account is geometric: it explains why the construction succeeds for some readout pairs and not others, and identifies a fitting choice that moves pairs across that boundary. Which *named* criteria will be separable lies outside it; the path is not fitted on the criteria used in the main experiment, and none of its quantities enters T.

S PER-CRITERION VALIDITY DOES NOT IMPLY AGGREGATE VALIDITY

Multi-objective reward models combine attribute scores by weighted summation (Wang et al., 2024a), whereas Section 4.5 averages the agreement endpoints of eleven separately evaluated criteria. This appendix asks a prior question about such aggregates: if every criterion in a group passes the validity screen on its own, does the group's aggregated readout also pass?

Define the aggregate readout of a group G as the normalised mean of its members' unit-normalised coefficient vectors, $w_G = \mathrm{normalize}\big(\mathrm{mean}_{d \in G}\, W_d/\|W_d\|\big)$, and let $\mathrm{acc}_{\mathrm{own}}(d)$ and $\mathrm{acc}_{\mathrm{agg}}(d)$ be the pairwise accuracies on criterion d of, respectively, d's own readout and the aggregate of the group containing d. Accuracies use the same convention as elsewhere: same-question response pairs with unequal labels, ties scored 0.5, question-clustered bootstrap.

Provenance. The three predictions were frozen on 2026-09-10, before EVAL was unblinded, and evaluated on 2026-09-16. Aggregate readout accuracies had not been computed previously, and this analysis was carried out after the main training result was known. These are criteria frozen before unblinding and evaluated later, on an endpoint distinct from Δ_B.

Aggregation selectively attenuates conciseness. Table 24 reports the ten criteria that pass the validity screen; `harmlessness` is excluded from this labelled-score validity analysis for having few non-tied evaluation pairs. The first prediction required the aggregation loss on `conciseness`

Table 24: Own versus aggregated readout accuracy on EVAL, Qwen2.5-1.5B. The aggregation loss is $\mathrm{acc}_{\mathrm{own}} - \mathrm{acc}_{\mathrm{agg}}$. One criterion loses substantially; for the rest the aggregate is within a few points of the criterion's own readout, and for five it is higher.

criterion	pairs	questions	$\mathrm{acc}_{\mathrm{own}}$	$\mathrm{acc}_{\mathrm{agg}}$	length	loss
`conciseness`	5834	105	0.653 [0.622, 0.685]	**0.516** [0.482, 0.551]	0.317	**+0.137**
`completeness`	10067	140	0.742 [0.715, 0.768]	0.728 [0.701, 0.752]	0.743	+0.014
`insightfulness`	6554	88	0.739 [0.715, 0.764]	0.725 [0.701, 0.749]	0.727	+0.014
`commonsense understanding`	15058	244	0.605 [0.580, 0.629]	0.602 [0.577, 0.626]	0.580	+0.003
`logical correctness`	8637	161	0.630 [0.598, 0.660]	0.630 [0.595, 0.661]	0.529	+0.000
`logical efficiency`	3275	46	0.681 [0.626, 0.737]	0.685 [0.633, 0.733]	0.509	−0.003
`factuality`	13160	225	0.616 [0.593, 0.640]	0.621 [0.596, 0.645]	0.597	−0.005
`logical robustness`	4199	58	0.706 [0.659, 0.749]	0.713 [0.668, 0.755]	0.548	−0.007
`comprehension`	26273	392	0.645 [0.626, 0.663]	0.666 [0.649, 0.682]	0.642	−0.021
`readability`	5132	113	0.590 [0.556, 0.623]	0.649 [0.618, 0.679]	0.550	−0.059

Table 25: Checks on generated text. Top: an independent LLM judge scores responses generated by the two arms on the FLASK rubric (1–5); entries are control minus treated with 95% question-bootstrap intervals over 21 and 13 questions. Bottom: an external reward model's excess win rate over 0.5, divided by the driving probe's, on 300 held-out prompts in an earlier construction; 1 means the reward model tracks the probe (Appendices T and U).

evaluator	condition	driver A	checker B
LLM judge	pair 1, control − treated	0.29 [0.00, 0.53]	0.33 [0.10, 0.57]
LLM judge	pair 3, control − treated	0.54 [0.15, 1.00]	0.62 [0.31, 1.00]
		ratio to driving probe	
ArmoRM	Qwen2.5-1.5B, random-direction control arm	0.95	
ArmoRM	Qwen2.5-1.5B, treated arm at matched KL	0.16	
ArmoRM	SmolLM2-1.7B, treated arm at matched KL	1.31	

to be at least 0.10 with the aggregate's interval upper bound below 0.60. Observed: loss +0.137, aggregate 0.516 [0.482, 0.551], an interval covering chance, while `conciseness`'s own readout reaches 0.653 [0.622, 0.685]. At the second model scale the same prediction gives loss +0.123 and upper bound 0.561. Both satisfy the pre-registered thresholds.

The second prediction was the control that makes the first interpretable: every qualifying criterion in the *other* group must retain an aggregate accuracy whose interval lower bound exceeds 0.5. All five do, at both scales. Averaging unit-normalised readouts therefore does not dilute signal in general; the failure is confined to one criterion.

Length-baseline prediction. We had also pre-registered an explanation: that aggregation loss tracks how far a criterion's length baseline departs from its group's mean, with a required rank correlation above +0.5. Observed: +0.467 at 1.5B and +0.442 at 0.5B over ten criteria. Neither reaches the threshold, and we record the mechanism as refuted. Following the pre-registration, we give no mechanistic account of the failure here; why an aggregate can lose a criterion that its own readout measures well remains open.

What this contributes. A scorecard assembled from individually validated criteria is not thereby a validated scorecard. This is a different route to the concern of Section 4.5: there, an aggregate loses sight of a training-induced change; here, an aggregate loses a criterion's signal on ordinary responses, before any training. The stopping rule fixed in advance permitted two models and no more; we did not vary the aggregation definition or the grouping after seeing these results.

T AN INDEPENDENT JUDGE SCORING NEWLY GENERATED TEXT

The main experiment measures agreement between policy preference changes and fixed criterion readouts. This appendix instead evaluates generated text: both arms generate responses, and an independent language-model judge scores them against the original FLASK rubric for the two criteria of the pair. Nothing about the measurement passes through a probe or a reward model.

Setup. The question subset and decoding were fixed before generation; the judge, rubric and cost ceiling were specified before scoring. For pairs 1 and 3 we take the questions of the original EVAL population annotated as applicable to *both* criteria of the pair: 21 questions for pair 1 and 13 for pair 3. Base, control and treated arms generate greedily at 512 tokens using the training chat template, giving 102 responses. The judge is `deepseek-v4-pro` at temperature 0, non-thinking, with the rubric and reference answer supplied as in the original annotation; all valid responses carry one system fingerprint. Eleven hidden re-scored items were interleaved with the 102: across their 33 criterion scores the judge reproduced its own score exactly 72.7% of the time with a mean absolute difference of 0.303 on the 1–5 scale. Every generated response is scored and retained, including repetitions and length-capped outputs; none is filtered. Intervals are question bootstraps with 2000 resamples.

The arms differ under a judge that never sees the readout.

change in mean judge score	pair 1 (21 q)	pair 3 (13 q)
A: control − base	$0.524\ [0.238, 0.810]$	$1.769\ [1.000, 2.692]$
A: treated − base	$0.238\ [0.000, 0.476]$	$1.231\ [0.308, 2.233]$
B: control − base	$0.619\ [0.381, 0.810]$	$1.692\ [0.923, 2.538]$
B: treated − base	$0.286\ [0.095, 0.476]$	$1.077\ [0.231, 2.154]$
A: control − treated	$0.286\ [0.000, 0.525]$	$0.538\ [0.154, 1.000]$
B: control − treated	$0.333\ [0.095, 0.571]$	$0.615\ [0.308, 1.000]$

On the bypassed criterion the two arms separate in both pairs, by 0.33 and 0.62 rubric points, with intervals excluding zero. The intervention is therefore visible to an evaluator that shares nothing with the instrument used to build it — no representation, no fitted coefficients, no reward model.

What this round is and is not designed to settle. The driving criterion moves with the bypassed one: both arms improve on both criteria relative to base, and the control arm improves more on both. The difference of the two contrasts is $+0.048\ [-0.238, +0.381]$ for pair 1 and $+0.077\ [-0.154, +0.308]$ for pair 3, so on these subsets the judge scores do not by themselves isolate a checker-specific effect. These estimates concern 21 and 13 questions selected before generation for joint criterion applicability. The raw rubric-point contrast measures behavioural score changes; it is distinct from the readout agreement endpoint used in the matched-KL comparison. This round measures behavioural changes independently of the readouts. Selective detection is established by the matched-KL readout experiments. The bypassed attribute has not stopped improving: the treated arm gains on it relative to base in both pairs. These numbers enter no aggregate in this paper.

T.1 SAME-RESPONSE ALIGNMENT BETWEEN DIRECT RATINGS AND FROZEN READOUTS

Connecting the two measurements. On the same generated responses for pair 1, direct ratings and eligible frozen readouts agree on the direction of all nine arm-level mean contrasts below. This follow-up links direct text assessment to representation-based scores by matching question and arm identities, making aggregate correspondence and within-question agreement separately measurable.

A distinct follow-up batch. This batch uses seed 20260922, pair 1 (21 questions) and pair 4 (32 questions) from the previously examined EVAL population. It is distinct from the 102-response pair-1/pair-3 batch above. Base, control and treated outputs give 159 responses, generated greedily with a 512-token cap. Pair-1 control/treated checkpoints are steps 1600/3000 (KL approximately $0.1764/0.1665$); pair 4 uses step 400 for both $(0.0570/0.0561)$. Checkpoint selection is developmental. The direct judge is `deepseek-v4-pro`; 16 hidden repeat items are excluded from effect estimates. All 159 response identities and source hashes were verified. Frozen linear readouts are

included; MLP eligibility follows the preceding SELECT validity gate, which admits insightfulness. No readout is selected using these generated-response outcomes.

Matched comparisons. For each question, we compare judge-score and readout-score differences for control minus base (C–base), treated minus base (T–base), and treated minus control (T–C). Direction agreement uses nonzero judge differences; judge ties are counted separately. Spearman correlations use the paired differences across all 21 questions, with 2,000 question-bootstrap resamples, seed 20260923, and pointwise 95% intervals. The two score scales retain their native units; their magnitudes are not equated.

Criterion	Readout	Contrast	Judge $+/-/0$	Same sign	ρ	95% CI
Completeness	linear	C–base	16/0/5	16/16	$+0.338$	$[-0.112, +0.689]$
Completeness	linear	T–base	10/1/10	10/11	$+0.495$	$[+0.039, +0.836]$
Completeness	linear	T–C	0/10/11	7/10	$+0.007$	$[-0.439, +0.422]$
Insightfulness	linear	C–base	14/0/7	14/14	$+0.198$	$[-0.285, +0.612]$
Insightfulness	linear	T–base	8/1/12	8/9	$+0.534$	$[+0.130, +0.812]$
Insightfulness	linear	T–C	1/9/11	8/10	-0.079	$[-0.510, +0.419]$
Insightfulness	MLP	C–base	14/0/7	14/14	$+0.081$	$[-0.330, +0.470]$
Insightfulness	MLP	T–base	8/1/12	9/9	$+0.440$	$[-0.022, +0.782]$
Insightfulness	MLP	T–C	1/9/11	9/10	$+0.197$	$[-0.287, +0.626]$

Interpretation at the measured resolution. The mean shifts correspond across measurement families, while the table preserves the direction balance needed to interpret individual comparisons. For insightfulness T–C, nine of ten judge changes are decreases: predicting decrease throughout also yields $9/10$, the qualified MLP's count. The linear readout and MLP differ in correctness on one question. All three T–C correlation intervals include zero. Accordingly, these are descriptive cross-measurement correspondences on shared questions, not evidence of MLP superiority or a validated checker-selection rule. Judge ties on a discrete scale do not establish unchanged quality or define readout false alarms.

Scope across both pairs. Pair 4 has no judge-resolved T–C changes: all logical-correctness scores are 1, and both trained arms score 1 on commonsense understanding (the base has one score of 2). No conditional direction rate is defined for those zero-event contrasts. All outputs, including repetition and truncation, remain included. The batch establishes an explicit same-response comparison; its developmental selection and shared EVAL questions distinguish it from independent confirmation of selective behavioural change or reduced missed detections.

U AN EXTERNAL REWARD MODEL SCORING NEWLY GENERATED TEXT

Every measurement in the main text reads a frozen linear direction out of hidden states, and the external reward experiment of Appendix F fits a deployed reward model back into that same representation. This appendix reports a pre-registered head-to-head that leaves the readout entirely: two arms *generate* responses, and an external reward model scores the text. The round was run on an earlier construction — one driving criterion and one removal, on a separate checkpoint family — and is therefore reported as independent evidence about generation, not as a sixth pair; its numbers enter no aggregate in this paper.

Setup. The prompts are 300 HelpSteer2 instructions that appear nowhere else in the project: not in any training set and not in any measurement partition. All arms generate under identical settings (temperature 0.7, top-$p = 0.9$, 256 new tokens, seed 0). For each prompt we form the pair (base response, arm response) with randomised presentation order and record length-controlled win rates under two evaluators: the probe direction that drives training, and ArmoRM (Wang et al., 2024a), an 8B multi-objective reward model with its own attribute heads. The comparison of interest is between two arms matched on conditional KL; the criteria and the prediction table were frozen before any generation.

Qwen2.5-1.5B. At matched displacement (KL 0.678 versus 0.656), the driving probe separates the two arms at a length-controlled win rate of 79.5% ($p = 1.9 \times 10^{-20}$), while ArmoRM reaches 54.6%

Table 26: Generation-level head-to-head on Qwen2.5-1.5B, earlier construction. Normalised ratio is ArmoRM's excess win rate over 0.5 divided by the driving probe's excess over 0.5, on the same 300 held-out prompts and the same generated responses. The random arm sets the exchange rate between the two evaluators.

arm compared against `audit_A@3000`	normalised ratio
`random@3000` (control)	0.951
`Bonly@3000`	0.846
`Alam50@3000`	0.168
`Aonly@1200` (matched KL)	$\mathbf{0.156}$

($p = 0.024$), below the pre-registered threshold of 0.01. The failure is not one of sensitivity in general: against the base model ArmoRM scores the two arms at 94.0% and 91.7%, so it registers that each arm improved. The between-arm comparison does not meet the pre-registered 0.01 significance threshold.

Table 26 normalises this. For each arm we divide ArmoRM's excess win rate over 0.5 by the driving probe's excess over the same reference, so a value near 1 means the external model tracks the probe and a value near 0 means it does not.

The control arm fixes the exchange rate at 0.951: on ordinary training displacement the two evaluators move together. The constructed arm sits at 0.156, six times lower, on real generations scored by a reward model that shares no parameters, no training data partition, and no functional form with the direction used to build it.

A second model family. On SmolLM2-1.7B the readout-level ordering reproduces: the ratio of retained agreement on the removed criterion to that on the driving criterion is 0.885 for `Bonly`, 0.659 for `audit_A` and 0.448 for `Aonly`, the same ordering as on Qwen. At the generation level the picture is different. At matched KL (0.317 versus 0.315) the driving probe separates the arms at 70.6% and ArmoRM at 76.9%, a normalised ratio of 1.306: here the external model orders the arms at least as sharply as the direction that produced them. The comparison has closely matched KL, and the base model has lower repetition (3-gram repetition 0.092 against 0.419).

The same nominal intervention is nearly invisible to this evaluator in one model family and fully visible in another, while the readout-level ordering holds in both. Visibility at the generation layer is therefore a property of the evaluator–direction pair as realised in a particular model, not a property of the intervention; this is the position of Section 1 applied to a case we did not construct for it, and it is the reason the census of Appendix N asks where the construction exists rather than assuming it travels.

What this establishes. Where the construction is available and the evaluator is fixed, an external reward model scoring genuine generations on unused prompts can be six times less sensitive to a matched-displacement change than the direction driving that change, with a same-round control at 0.951 showing the two evaluators otherwise agree. The generation-level gap is demonstrated in one model family; the readout-level ordering that accompanies it is demonstrated in two. ArmoRM is an independent evaluator rather than an instance of the criterion being removed, and its win rate is reported as an external model's judgement of the generated text, not as a measurement of that criterion.

V EQUAL-LENGTH PLACEBO ARMS

This appendix gives the construction, the frozen quantities, and the per-pair diagnostics behind Section 4.3.

Construction. In the measurement basis S, with $a = \mathrm{norm}(P_S W_A)$ and $\tilde{b}$ the sign-aligned $\mathrm{norm}(P_S W_B)$ and $c = a^\top \tilde{b}$, the treated direction is $u_{\mathrm{real}} = \mathrm{norm}(a - c\,\tilde{b})$. The placebo draws $r = c\,a + \sqrt{1-c^2}\,v$ with $v \perp a$, $\|v\| = 1$, $v \in S$, and sets $u_{\mathrm{pl}} = \mathrm{norm}(a - c\,r)$. Both $\tilde{b}$ and r satisfy $a^\top(\cdot) = c$, so the removed length is c, the residual norm is $\sqrt{1-c^2}$ and the rotation angle is $\arcsin c$ in both arms. These are algebraic identities of the construction, not empir-

Table 27: Equal-length placebo arms. The placebo removes the same length at the same rotation angle in a different direction drawn from the same equal-angle set; pair indices are those of Table 1. Paired differences use identical qid-clustered resamples ($B = 2000$). Pairs 4 and 5 admit no matched unit.

pair	Δ_B directional	Δ_B placebo	paired difference
1	$+0.0889$ $[+0.0809, +0.0962]$	$+0.0088$ $[+0.0032, +0.0148]$	$\mathbf{+0.0801}$ $[+0.0711, +0.0885]$
2	$+0.0387$ $[+0.0332, +0.0445]$	$+0.0031$ $[-0.0031, +0.0095]$	$\mathbf{+0.0357}$ $[+0.0280, +0.0428]$
3	$+0.0754$ $[+0.0683, +0.0824]$	$+0.0282$ $[+0.0190, +0.0373]$	$\mathbf{+0.0472}$ $[+0.0352, +0.0591]$

Table 28: Quantities fixed before any placebo arm was trained. η_B is the residual scoring sensitivity of the removed criterion under each arm's construction. The constraint $a^\top r = c$ matches the removal geometry; the last column records the realised overlap with the bypassed direction.

pair	c	δ	η_B real	η_B placebo	$r^\top \tilde{b}$
1 completeness → insightfulness	0.5602	34.07°	0.1810	0.8988	$+0.2551$
2 logical efficiency → conciseness	0.3590	21.04°	0.2413	1.0591	-0.0239
3 logical robustness → comprehension	0.5837	35.71°	0.1459	0.8510	$+0.2723$
4 logical correctness → commonsense	0.6209	38.38°	0.1974	0.8070	$+0.3685$
5 readability → factuality	0.4352	25.80°	0.0271	0.9372	$+0.1364$

ical findings. The vectors v come from a fixed seed shared across the five pairs, with a guard $|r^\top \tilde{b}| \leq c^2 + 0.05$ and redraw on violation.

Table 28 lists these. A placebo η_B above one, as in pair 2, is not an anomaly and is not read as the placebo making B more responsive: η is a ratio of sensitivities whose numerator and denominator are both base-model geometric quantities.

Which pairs yield a unit, and why the other two do not. The KL target, tolerances and arm-gap limit are those of the main experiment. Pair 4 fails on the arm gap: the control arm's closest checkpoint sits at KL 0.3479 and the placebo arm's at 0.4993, a gap of 0.1514 against a limit of 0.10. Pair 5 fails on the control arm itself, whose trajectory reaches only 0.2376, an absolute deviation of 0.2124 against a limit of 0.15. Both are left blank. Three of five pairs therefore yield a matched unit. The driving-side retention ratios on the placebo arms are 0.979, 0.990 and 0.991 for pairs 1, 2 and 3, all above the pre-registered admission threshold.

Extension, and a second measurement of run-to-run variation. The schedule was frozen at 3000 steps with a single pre-registered extension to 13000 for any arm whose KL trajectory peaked below 0.45, decided on the KL trace alone. Pairs 1 and 5 triggered it, at peaks of 0.3388 and 0.2468; pairs 2, 3 and 4 peaked at 0.7988, 0.5453 and 0.5591 and did not. The extension was executed as a fresh run under the same seed and hyper-parameters rather than as a resumption, which yields a second reading on the variation floor of Appendix H: at shared steps the two runs of pair 1 differ by up to 0.1045 in conditional KL (0.3028 against 0.2043 at step 1600, a relative difference of 32.5%), and those of pair 5 by up to 0.0396. This is larger than the same-seed range reported there and concerns two runs of one arm rather than repetitions of one run. Because the two runs are not the same trajectory, we fixed before reading any agreement statistic that the extension run is the run of record for a pair that triggered the extension, and that checkpoints from the two runs are never mixed within a comparison.

Scope of the reported quantities. The pre-registration excludes placebo arms from T, and T is defined over five pairs; no aggregate over the three placebo units appears in this paper, and the per-pair paired differences of Table 27 are the reported result. The blanks in pairs 4 and 5 are a failure to obtain a matched unit and carry no information about the size of the effect in those pairs.

Registered directional test. The pre-specified criterion required the paired difference to exclude zero before attributing the separation to the direction of removal. All three matched pairs satisfy it. The separate zero-effect prediction for the placebo holds in pair 2 but not in pairs 1 and 3. The

Table 29: Response of each criterion to optimisation along a reward model's own readout, against a KL-matched random direction. Agreement is length-adjusted; the difference and its interval are computed on identical qid-clustered resamples ($B = 2000$). Criteria are ordered by overlap with the reward direction.

criterion	overlap	$\alpha^{\perp}$ reward	$\alpha^{\perp}$ random	difference
comprehension	0.4758	0.6144	0.5219	$+0.0925$ $[+0.0820, +0.1034]$
completeness	0.4655	0.6342	0.5481	$+0.0862$ $[+0.0762, +0.0963]$
insightfulness	0.4375	0.6300	0.5526	$+0.0774$ $[+0.0671, +0.0878]$
factuality	0.3799	0.5992	0.5299	$+0.0693$ $[+0.0594, +0.0794]$
commonsense understanding	0.3798	0.6075	0.5183	$+0.0892$ $[+0.0789, +0.0998]$
readability	0.3733	0.5724	0.4888	$+0.0837$ $[+0.0737, +0.0944]$
logical robustness	0.3619	0.5773	0.5254	$+0.0519$ $[+0.0418, +0.0623]$
logical efficiency	0.3066	0.5695	0.5104	$+0.0591$ $[+0.0487, +0.0698]$
logical correctness	0.3064	0.5575	0.4872	$+0.0702$ $[+0.0588, +0.0814]$
conciseness	0.2087	0.5667	0.4939	$+0.0728$ $[+0.0618, +0.0839]$
harmlessness	0.0473	0.5105	0.5070	$+0.0036$ $[-0.0069, +0.0144]$

paired contrast therefore measures the additional effect of directed removal over the nonzero placebo response, rather than establishing that equal-length removal has no effect.

W ORDINARY REWARD-MODEL OPTIMISATION: PER-CRITERION DETAIL

This appendix tests directional coverage under an unmodified reward readout (Section 4.2).

The two arms. Write u_{RM} for the ambient readout direction of the deployed reward model of Appendix F, and

$$u_{\mathrm{RM}}^{S} = \mathrm{norm}(P_S\, u_{\mathrm{RM}}) \tag{24}$$

for its normalised projection into the measurement subspace, provided $\|P_S u_{\mathrm{RM}}\| > 0$. The reward arm is trained along u_{RM}^{S}, with nothing removed. The control arm is trained along a random unit direction r drawn in the same subspace S under two guards fixed before drawing: $|r^{\top} u_{\mathrm{RM}}^{S}| \le 0.10$ and $\max_X |r^{\top} v_X| \le 0.20$ over all eleven criteria. The realised values are 0.0390 and 0.1246, the latter on `completeness`; the first draw satisfied both. The control direction file has the same structure as the reward arm's, so the two arms run through an identical training path rather than a separate code branch.

KL cells. Under the main rule — target 0.45, relative deviation ≤ 0.25, absolute ≤ 0.15, between-arm gap ≤ 0.10 — the reward arm is taken at step 2200 with KL 0.4109 and the control arm at step 5600 with KL 0.4328, a gap of 0.0219. The steps differ because the arms are matched on displacement rather than on optimiser steps, as elsewhere in this paper.

Positive control. The reward model's own attribute heads are detected in the reward arm: helpfulness 0.5454, correctness 0.5405, coherence 0.5341, and the combined head 0.5418, each with an interval above chance. An arm trained along a reward model's readout is therefore visible to that reward model, which is the minimum the design requires.

Ordering and zero-response predictions. Two predictions were fixed before the control arm existed: that the rank correlation between overlap and response difference would be positive, and that the three lowest-overlap criteria would each have a difference covering zero. The first holds at $\rho = +0.709$, $[+0.518, +0.855]$. The second does not: `logical correctness` and `conciseness` both exclude zero and only `harmlessness` covers it. We record the prediction as not holding and make no claim from the one criterion that matches it. `harmlessness` is also the criterion this work excludes elsewhere for having few non-tied evaluation pairs, so its null is not evidence of a region ordinary optimisation cannot reach.

Where the reward direction lives. The reward model's readout is fitted in the ambient hidden-state space, and the measurement subspace S used for the criterion readouts retains $\|P_S u_{\mathrm{RM}}\|/\|u_{\mathrm{RM}}\| = 0.443$ of its length, equivalently 19.7% of its squared norm. That ratio is also the cosine between u_{RM} and the trained direction u^S_{RM} of Equation (24). Both the overlap column of Table 29 and the guard on the random draw are computed inside S, against u^S_{RM}, so the comparison is internally consistent; measured against the ambient direction instead, the random control's overlap is 0.017 and the eleven criteria span 0.005 to 0.198.

X CHECKER SELECTION: RANKINGS AND PER-PAIR CELLS

This appendix gives the detail behind Section 5. The analysis is post hoc in the sense that the evaluation partition was already unblinded when it was designed; what was fixed in advance, and before any cell was computed, is the three ranking rules with their directions, the quantity compared, the judgement point $k = 1$, and the reading of each outcome. The three rankings were fixed before any detection value was read, and independently reproduced afterwards with identical results.

The rules. For a driving criterion A, candidates are the nine criteria excluding A and `harmlessness`, and including the targeted checker B: the agreement-based baseline ranks by $C_2(A, X)$, the share of same-question pairs on which the two readouts differ in the same direction, descending. The geometric rule ranks by $\eta^{\text{single}}_X = |u_X^\top (W_X/\sigma)|/|a^\top (W_X/\sigma)|$ descending, where $u_X = \mathrm{norm}(a - (a^\top v_X)v_X)$ is the direction built to evade X alone. The oracle reference ranks by $\mathrm{rem}_X = v_X^\top u_A - v_X^\top u_{A\setminus B}$ ascending; it uses $u_{A\setminus B}$ and is reported as a direction-informed reference, not as a bound on measured detection. The floor draws k of the nine uniformly without replacement, 2000 times.

Quantities. With $\alpha_X(\mathrm{treat})$ obtained as $\alpha_X(\mathrm{ctrl}) - \Delta_X$ on identical resamples, let $S_R(j, k)$ be the set of top-k criteria selected by rule R for pair j. Then $M_R(j, k) = \max_{X \in S_R(j,k)}(\alpha_X(\mathrm{treat}) - 0.5)$ is the primary quantity, taken as a maximum because one checker raising an alarm is enough in practice, and $V_R(j, k)$ counts the chosen criteria whose interval lower bound exceeds 0.5. Differences between rules are computed on the same resamples.

Candidate-pool power. A pair was to be set aside if more than half its candidates failed to detect the change in the control arm. None did: eight of the nine candidates detect in pair 3 and all nine in the other four pairs, so all five pairs are judged.

Primary comparison and additional checkers. At the fixed judgement point $k = 1$, the geometric rule minus the agreement baseline is -0.1483 $[-0.1582, -0.1383]$ on paired resamples; the oracle reference minus the geometric rule is $+0.1130$ $[+0.1048, +0.1211]$. No rule–pair cell at $k = 1$ exceeds the random floor's 95th percentile. At $k = 3$, all three agreement-selected criteria retain detection above the fixed reference in every pair. These results support checking absolute baseline responsiveness alongside relative sensitivity; they do not establish a selection rule superior to random choice.

What the geometric rule selects. It ranks `conciseness` first for every one of the five driving criteria, and `conciseness` retains -0.0424, $+0.0004$, -0.0519, $+0.0120$ and $+0.0086$ above the fixed 0.5 reference in the five treated arms. Its ranking quantity is a ratio of a criterion's sensitivities, so a criterion whose readout carries little sensitivity scores well on it without detecting anything. The direction-informed reference places B last in all five pairs: $\mathrm{rem}_B = c$ is the largest value in these candidate panels. This ordering is observed here, not forced for arbitrary additional readouts.

Baseline-aware selection as a prospective test. Combining residual sensitivity with baseline detection is motivated by the ratio behaviour above. It remains an untested proposal; the reported comparison retains the three originally fixed rules. Evaluating an amended rule on new pairs would test its predictive value separately from the outcomes that motivated it.

Y First-order mechanism: criteria, traces, and provenance

Order of operations. The predictions below were fixed before any gradient was accumulated, in a document that also fixed what each outcome would license. The predictions were computed, written to a file, and the file's SHA-256 recorded in that document (`d1a7ddba...f933e7d3b`) before the trained checkpoints were measured; the hash was re-verified against that record before the comparison was computed. The ε bisection reads conditional KL only, and its full trace is reported for each arm.

Feasibility: met. All ten arms produced a displacement whose induced KL fell inside the tolerance, against a requirement of four. The accumulated gradient norms lie between 103 and 119 and the chosen ε between 1.6×10^{-3} and 3.2×10^{-3}. Each displaced model differs from the base model by a mean absolute log-probability change between 0.045 and 0.130, far above the threshold below which an adapter is treated as inactive.

P2, sign across pairs: met. Reported in Section 4.5. This was the load-bearing criterion, and the only one whose targets had never been measured in any form when the predictions were fixed.

P1, ordering across criteria: not met. The pre-registered form required a positive rank correlation between predicted and measured per-criterion responses in every pair *and* in the pooled set. Pairs 1–4 give $+0.8818$ $[+0.5727, +0.9545]$, $+0.9364$ $[+0.8089, +0.9818]$, $+0.8818$ $[+0.5000, +0.9182]$ and $+0.9909$ $[+0.9636, +1.0000]$; pair 5 gives -0.4455 $[-0.8275, +0.1457]$; pooled, $+0.7533$ $[+0.6574, +0.8271]$. The conjunction is not supported. The intervals are post-hoc descriptive analyses; the original criterion concerned the signs of the correlations, and its verdict is retained.

P3, directional separation: not met. With the real treated arm as reference and the equal-length placebo arm as comparison, the quantity is $\alpha_B(\text{real}) - \alpha_B(\text{placebo})$, negative when the directed removal costs the checked criterion more than the undirected one does. Predicted and measured: pair 1, -0.1069 against -0.0201 $[-0.0295, -0.0107]$; pair 2, -0.0392 against $+0.0108$ $[-0.0047, +0.0262]$; pair 3, -0.0178 against -0.0070 $[-0.0120, -0.0020]$. Two of three agree in sign and the third's measured value covers zero, so this is not a significant reversal; the pre-registered form asked for three of three and is recorded as not met.

Raw-gradient prediction and adaptive training. The instrument uses a single accumulated gradient direction; the training experiments use AdamW with LoRA. Adaptive preconditioning can rotate and rescale the update, so the comparison tests raw-gradient first-order predictivity. Gradient clipping, present in the training loop, is not applied when accumulating, because clipping each batch and then summing is not the same operation as summing and then clipping, and the first-order object is the sum. The measured targets are checkpoints from runs trained on 2026-09-17 specifically for this section, under the main experiment's hyper-parameters and seed but as a separate run; conditional KL between two runs of one arm at a shared step has been observed to differ by up to 0.10 in this work, so no checkpoint from these runs is mixed with the main table.

Y.1 Local response under optimizer-direction substitutions

At fixed post-update states, changing the update direction changes the local criterion response under a common calibration target. We compare the current batch gradient, momentum, and their adaptively preconditioned directions at eight states (two pairs, two arms, and steps 100 and 400). Each direction is separately scaled to a KL target of 10^{-4} on eight fixed responses. Figure 10 reports all 48 response contrasts. For pair 4 at step 100, preconditioning momentum increases checker agreement by 0.0646 and 0.0716 in the control and treated arms; at step 400 the corresponding changes are -0.0064 and $+0.0068$. The response to an update therefore depends on both its direction and the state at which it is applied. These local substitutions complement the cumulative trajectory measurement. They do not identify the cause of its crossover. The states are developmental; intervals are post-hoc, paired question-bootstrap intervals (2,000 resamples), pointwise and conditional on the fixed states. Calibration responses overlap with EVAL, so the common target does not establish EVAL-wide matched KL. Directions exclude weight decay, and separately calibrated response contrasts are not an additive decomposition of AdamW.

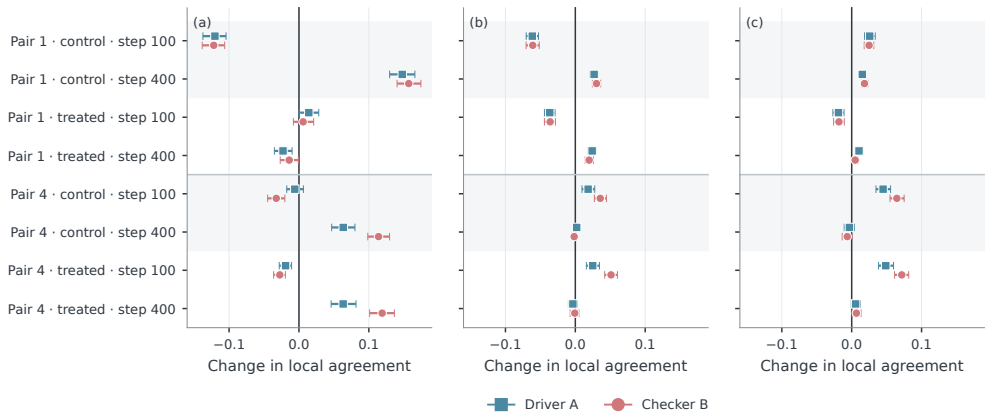

Figure 10: Local criterion response under optimizer-direction substitutions. Panels compare (a) momentum against the current gradient, (b) preconditioned gradient against the gradient, and (c) preconditioned momentum against momentum. All eight states are shown for both the driver and checker. Points are response differences and bars are pointwise 95% question-bootstrap intervals. Each direction is calibrated separately on the same eight responses; the comparison is a development diagnostic.

Y.2 Optimal residual direction and the update-to-response map

Optimality in coefficient space. Let $a, b \in S$ be unit vectors, $s = a^\top b$, and $c = |s| < 1$. Write $r = a - sb$. For any feasible $u \in S$ with $\|u\| = 1$ and $b^\top u = 0$, Cauchy–Schwarz gives $a^\top u = r^\top u \le \|r\|$. Since $\|r\| = \sqrt{1 - c^2} > 0$, equality holds uniquely at $u = r/\|r\|$:

$$u_{A\setminus B} = \operatorname*{arg\,max}_{u \in S,\, \|u\|=1,\, b^\top u = 0} a^\top u, \qquad \max a^\top u = \sqrt{1 - c^2}. \tag{25}$$

The angle to a is minimised by the same direction and equals $\arccos\sqrt{1 - c^2} = \arcsin c$. When $c = 0$, the optimum is a itself. When $c = 1$, the residual is zero and every feasible direction has zero alignment with a; if S is one-dimensional there is no feasible unit direction. These statements concern coefficient alignment, not optimal retention of a trained policy's detection signal.

From a policy update to a probability margin. For fixed responses y_i^+, y_i^- to question x_i, define $\ell_\theta(y \mid x) = |y|^{-1} \log p_\theta(y \mid x)$ using the response-token convention of the endpoint, and $m_i(\theta) = \ell_\theta(y_i^+ \mid x_i) - \ell_\theta(y_i^- \mid x_i)$. The superscripts index the two responses and need not denote preference. For a displacement d from θ, Taylor expansion gives

$$d_i := m_i(\theta + d) - m_i(\theta) = g_i^\top d + R_i, \qquad g_i = \nabla_\theta m_i(\theta). \tag{26}$$

If m_i is twice differentiable along the displacement and its Hessian operator norm is bounded by L_i on that segment, then $|R_i| \le L_i \|d\|^2/2$. Consequently, if $|g_i^\top d| > L_i \|d\|^2/2$, the linearised and actual changes have identical signs. Small displacement alone does not ensure sign preservation for a margin near zero.

Connection to the measured agreement. Let $q_{X,i} = s_X(h_i^+) - s_X(h_i^-)$ be the fixed criterion-readout margin. For nonzero margins, a response pair registers agreement exactly when $q_{X,i} d_i > 0$. Under the sign-preservation condition, this is equivalent to $q_{X,i} g_i^\top d > 0$. The reported endpoint retains its original tie rule; this non-tie identity neither changes that rule nor its chance reference. The base-to-checkpoint endpoint uses $\theta = \theta_0$ and the total policy displacement. Local substitutions instead use their specified anchor state. The linear approximation describes a local map and is not assumed accurate for an entire training trajectory.

Mechanistic interpretation and evidence. The constructed direction assigns preferences; optimisation maps those preferences to parameter updates; the model's margin gradients map updates to probability changes. Fixed criterion rankings measure detection of those changes. Thus coefficient

overlap supplies an intervention and a geometric profile, while the update and model state enter the realised response. The five-pair matched-KL comparison establishes endpoint separation. The base-model gradient instrument tests early sign predictivity, the independent seed tests temporal reproducibility, and the eight-state direction substitutions test local direction and state dependence. The Taylor identity provides the mathematical connection among these measurements; it is not an additional empirical prediction test or a decomposition of the full AdamW trajectory.