OBERON: DOES DETECTING REWARD HACKING INFORM ITS REMOVAL?

Anonymous authors
Paper under double-blind review

ABSTRACT

Reward models (RMs) are central to language model alignment, and interventions on their internal representations can mitigate reward biases. However, a direction that reliably detects a response property does not necessarily reduce its associated reward gap when erased. To address this challenge, we introduce OBERON, which jointly learns a single direction for detection and reward-gap attenuation in a frozen RM. We first formalize directional readout and representation erasure within an observation–intervention framework and show that detector scores alone do not generally identify signed erasure effects over a family of affine downstream maps. We then derive a pairwise upper bound on a population risk combining both objectives and use its empirical form to couple a paired logistic loss with a penalty on reward gaps remaining after erasure of the same direction. We evaluate five RMs on fixed response pairs involving formatting, length, sycophancy, and conflicts between quality and style. In each condition, OBERON achieves the highest average proportion of pairs with both correct paired detection and reduced absolute reward gaps among the compared methods, with an average drop in preference accuracy below one percentage point. On formatting and length, the joint success rates averaged across the five RMs reach 90.5% and 70.3%, compared with 83.5% for score penalty and 64.0% for difference of means, respectively.

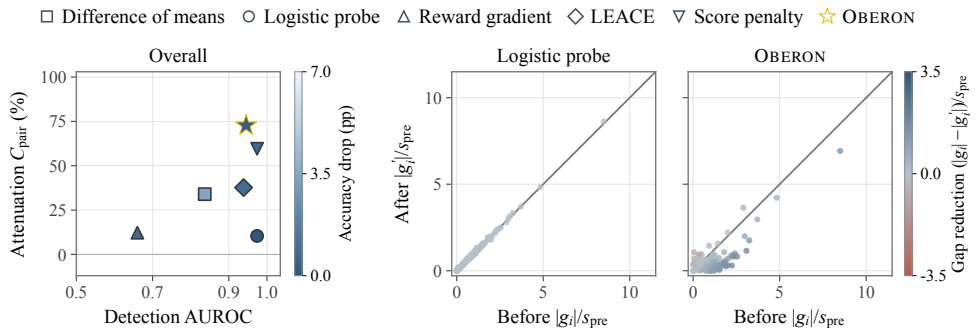
1 INTRODUCTION

Reinforcement learning from human feedback (RLHF) is widely used to align large language models (LLMs) with human preferences (Ouyang et al., 2022; Lambert, 2025). In RLHF, a reward model trained on human preference comparisons provides the reward signal for subsequent policy optimization. Despite its empirical effectiveness, RLHF remains vulnerable to reward hacking, as policy optimization can exploit discrepancies between reward-model predictions and the intended objective, increasing predicted reward without improving response quality (Gao et al., 2023; Skalse et al., 2022). Reward models exhibit potentially exploitable biases, including sensitivity to response length and formatting (Liu et al., 2025b; Zhang et al., 2025), as well as a preference for responses that agree with users even when their stated beliefs are incorrect (Sharma et al., 2024).

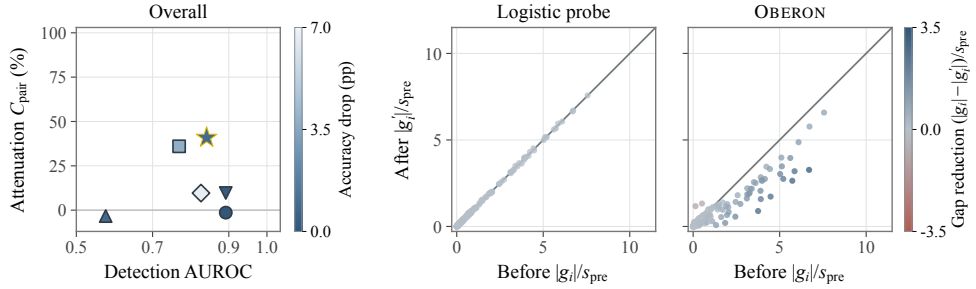
Existing approaches seek to mitigate reward hacking by improving reward estimation or regulating policy optimization. Representative methods disentangle quality from length (Chen et al., 2024), augment preference data (Liu et al., 2025a), use reward ensembles for conservative optimization (Coste et al., 2024), or regularize preference optimization with a supervised fine-tuning loss (Liu et al., 2024b). A complementary line of work targets specific bias-related signals within already-trained reward models. These methods use internal representations to identify such signals and guide corrections to reward predictions. Linear-probe penalties adjust output rewards using learned sycophancy scores (Papadatos & Freedman, 2024), while directional interventions modify identified components through representation projection (Fein et al., 2026) or reward-head editing (Liu et al., 2026).

However, detecting a bias-related signal does not by itself establish the reward effect of erasing the corresponding representation component. Prior work distinguishes recoverable information from behavioral influence (Elazar et al., 2021) and evaluates detection and steering as distinct capabilities (Wu et al., 2025). In the left panels of Figure 1, which report five-model averages, the logistic probe (Alain & Bengio, 2016) achieves high AUROC, yet projecting out its direction

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.



(a) Formatting (`fmt`)



(b) Length (`len`)

Figure 1: **Detection, reward-gap attenuation, and collateral quality.** Left panels report detection AUROC and reward-gap attenuation averaged across five reward models, with color indicating mean preference-accuracy drop (pp). Right panels show illustrative single-fit results for logistic-probe projection and OBERON on one reward model. Pre- and post-intervention absolute gaps are normalized by the mean absolute baseline gap s_{pre} within each condition. Color indicates normalized gap reduction, and points below the diagonal indicate attenuation.

yields limited reward-gap attenuation on formatting pairs and slight amplification on length pairs. LEACE (Belrose et al., 2023), using its native erasure operator, achieves greater attenuation than logistic-probe projection in both conditions, but with lower AUROC and larger preference-accuracy drops. These comparisons illustrate that detection performance alone does not characterize reward-gap attenuation or collateral quality. This motivates incorporating the reward effect of erasure into direction learning, while evaluating collateral quality separately.

To this end, we introduce OBERON, which jointly learns a single direction for detection and reward-gap attenuation in a frozen reward model. We first formalize these roles as directional readout and representation erasure, showing that detector-score information alone does not generally identify the signed erasure effect within a specified family of affine downstream maps. We then define a population risk capturing both objectives and derive a pairwise upper bound. We use an empirical version of this bound as the OBERON training objective, combining a paired detection loss with a penalty on reward gaps remaining after erasure of the same direction. We evaluate five reward models on formatting, length, sycophancy, and conflicts between quality and style. In each condition, OBERON achieves the highest model-averaged joint detection-and-reduction rate among the compared methods, with an average preference-accuracy drop below one percentage point. These results concern fixed response pairs rather than robustness under subsequent policy optimization.

Our main contributions are summarized as follows.

- **Theoretically,** we formulate the observation–intervention framework to distinguish directional readout from representation erasure and establish that detection information alone does not generally determine the signed erasure effect under affine downstream maps.
- **Algorithmically,** we derive a pairwise surrogate from a population-risk upper bound and introduce OBERON, which learns a single direction by combining a paired detection loss with a penalty on post-erasure reward residuals.

- **Experimentally,** we evaluate detection, reward-gap attenuation, and collateral quality across five reward models and four conditions. OBERON achieves the highest average joint detection-and-reduction rate in every condition while largely preserving preference accuracy.

2 RELATED WORK

Reward learning and policy optimization. Reward hacking occurs when optimization exploits discrepancies between a learned reward and the intended objective (Skalse et al., 2022; Gao et al., 2023). Existing mitigation methods modify reward learning or regulate subsequent policy optimization. ODIN disentangles quality from length (Chen et al., 2024), RRM augments preference data through prompt–response recombination (Liu et al., 2025a), and CausalRM learns factorized reward representations (Yang et al., 2026). Other approaches use reward ensembles for conservative optimization (Coste et al., 2024) or regularize preference optimization with a supervised fine-tuning loss (Liu et al., 2024b). These approaches improve reward learning or limit its exploitation during policy optimization, but do not directly address targeted correction within a trained reward model. OBERON considers this complementary setting by learning a representation direction that supports both detection and post-hoc reward-gap attenuation.

Representation-based reward-model interventions. Linear probes assess whether information is recoverable from representations (Alain & Bengio, 2016), while control tasks examine whether probe performance reflects the representation rather than probe capacity (Hewitt & Liang, 2019). INLP (Ravfogel et al., 2020) and LEACE (Belrose et al., 2023) remove linearly recoverable information, while amnesic probing evaluates the resulting behavioral changes (Elazar et al., 2021). Related work further distinguishes detection from steering (Wu et al., 2025; Galeone et al., 2026). In reward models, probe-based penalties adjust output rewards (Papadatos & Freedman, 2024), activation editing modifies bias-associated neurons (Shinoda et al., 2026), representation projection removes identified bias directions (Fein et al., 2026), and reward-head editing suppresses identified hacking subspaces (Liu et al., 2026). These methods demonstrate that internal signals can guide post-hoc reward correction, but the detectability of a direction does not by itself characterize its reward effect under erasure. OBERON addresses this mismatch by learning the same direction from both paired detection and the reward residual produced by its erasure.

3 THE OBSERVATION–INTERVENTION FRAMEWORK

3.1 PROBLEM FORMULATION

Observation and intervention. For a given representation direction, we distinguish an observational contrast from an interventional effect. The former is the signed difference between the hack and clean conditional mean scores along that direction. The latter is the change in the model's hack–clean mean reward gap induced by erasing the corresponding representation component. The interventional effect additionally depends on the frozen downstream computation. Our central theoretical question is whether this effect is identifiable from the joint distribution of directional scores and condition labels alone, without reevaluating the frozen model under representation erasure.

Setup and notation. Let (X, Y, Z) be sampled from the audit population, where X is a prompt, Y is a response, and $Z \in \{0, 1\}$ indicates the condition, with $Z = 1$ for hack and $Z = 0$ for clean. For each prompt, the audit population contains one matched pair consisting of a hack response and a clean response, and all pairs receive equal weight. By construction, the conditional distribution of X is identical under the two conditions. A frozen reward model assigns the scalar reward $R(X, Y)$. At a fixed internal layer ℓ, let $H := h_\ell(X, Y) \in \mathbb{R}^d$ denote the hidden representation and factor the reward computation as $R(X, Y) = G(H)$, where $G : \mathbb{R}^d \to \mathbb{R}$ is the frozen downstream map. Let P denote the induced joint distribution of (H, Z). We assume $\Pr(Z = z) > 0$ for $z \in \{0, 1\}$ and that all expectations below are finite. The hack–clean conditional mean displacement is

$$\delta_H := \mathbb{E}[H \mid Z = 1] - \mathbb{E}[H \mid Z = 0]. \tag{1}$$

The subtraction orients δ_H from the clean conditional mean representation to the hack conditional mean representation. For any unit direction $v \in \mathbb{R}^d$, define the linear score and orthogonal erasure

operator by

$$s_v(h) := v^\top h, \qquad Q_v := I - vv^\top, \qquad Q_v h = h - v(v^\top h), \tag{2}$$

where I is the $d \times d$ identity matrix. Since $\|v\|_2 = 1$, Q_v is the orthogonal projector onto $v^\perp$ and removes the component of h along v. The observational branch leaves H unchanged and uses only the joint distribution of $(s_v(H), Z)$. The interventional branch instead replaces H with $Q_v H$ and evaluates $G(Q_v H)$ while keeping G and all model parameters fixed.

3.2 Formalizing Observation and Intervention

Definition 1 (Observation–intervention functional). *Let $F : \mathbb{R}^d \to \mathbb{R}$ be a scalar readout and let $T : \mathbb{R}^d \to \mathbb{R}^d$ be a representation transformation. Define*

$$\mathcal{J}(F, T) := \mathbb{E}[F(T(H)) \mid Z = 1] - \mathbb{E}[F(T(H)) \mid Z = 0]. \tag{3}$$

The functional $\mathcal{J}$ maps each readout–transformation pair (F, T) to the class-conditional mean difference of $F(T(H))$. Observation fixes $T = I$ and evaluates the readout F, whereas the effect of an intervention T is measured by fixing $F = G$ and comparing $\mathcal{J}(G, T)$ with the baseline $\mathcal{J}(G, I)$.

Observation through a linear readout. For a unit candidate direction v, the observational branch leaves the representation unchanged and applies the linear readout $s_v(h) = v^\top h$:

$$D(v) := \mathcal{J}(s_v, I) = \mathbb{E}[s_v(H) \mid Z = 1] - \mathbb{E}[s_v(H) \mid Z = 0] = v^\top \delta_H. \tag{4}$$

The final equality follows from linearity of conditional expectation and the definition of δ_H. Thus, $D(v)$ is the oriented hack–clean conditional mean-score contrast along v, with $D(-v) = -D(v)$. It captures only first-moment separation and is not a general measure of detection performance.

Intervention through representation erasure. For the same direction v, the interventional branch keeps the frozen downstream map G fixed and replaces H with $Q_v H$. Define the baseline hack–clean mean reward gap Δ_0 and its post-erasure counterpart Δ_v as

$$\Delta_0 := \mathcal{J}(G, I), \qquad \Delta_v := \mathcal{J}(G, Q_v). \tag{5}$$

Definition 2 (Signed removal ratio). *Suppose $\Delta_0 \neq 0$. For each unit direction v, define*

$$C(v) := 1 - \frac{\Delta_v}{\Delta_0}. \tag{6}$$

The signed ratio $C(v)$ is not restricted to $[0, 1]$, with $C(v) = 0$ corresponding to an unchanged mean reward gap ($\Delta_v = \Delta_0$) and $C(v) = 1$ to a zero post-erasure mean gap ($\Delta_v = 0$). Values in $(0, 1)$ indicate attenuation without sign reversal, whereas negative values indicate amplification with the same sign and values above one indicate sign reversal. Unlike $D(v)$, $C(v)$ is orientation-invariant because $Q_{-v} = Q_v$ implies $C(-v) = C(v)$. Throughout, *intervention* denotes the computational transformation Q_v, not a structural-causal intervention on an underlying semantic concept.

3.3 Observational Scores Do Not Generally Identify the Erasure Effect

The contrast $D(v)$ captures only the difference between class-conditional score means. We ask whether the full joint law of $(s_v(H), Z)$ identifies $C(v)$. Under an affine downstream map, the following results show that removal also depends on a downstream directional coefficient that this law does not generally determine.

Lemma 1 (Factorization under an affine downstream map). *Let $G(h) = c + w^\top h$, where $c \in \mathbb{R}$ and $w \in \mathbb{R}^d$. For any unit direction v, we have $\Delta_0 = w^\top \delta_H$ and $\Delta_0 - \Delta_v = D(v)(w^\top v)$. Whenever $\Delta_0 \neq 0$, the signed removal ratio factorizes as follows (proof in Appendix A.1):*

$$C(v) = \frac{D(v)(w^\top v)}{\Delta_0} = \frac{(v^\top \delta_H)(w^\top v)}{w^\top \delta_H}. \tag{7}$$

Proposition 1 (Non-identifiability of removal from observational scores). *Fix the joint law P of (H, Z), a unit direction v satisfying $0 < |v^\top \delta_H| < \|\delta_H\|_2$, and a scalar $\beta \neq 0$. Let $\mathcal{G}_\beta$ consist of affine maps $G_{c,w}(h) = c + w^\top h$, where $c \in \mathbb{R}$ and $w \in \mathbb{R}^d$ satisfies $w^\top \delta_H = \beta$, with no restriction on $\|w\|_2$. The signed removal ratio $C_G(v)$ is not point-identifiable over $\mathcal{G}_\beta$ from the joint law of $(s_v(H), Z)$ and the baseline gap. All maps in $\mathcal{G}_\beta$ share this law and satisfy $\mathcal{J}(G, I) = \beta$, yet for every $\gamma \in \mathbb{R}$ there exists $G_\gamma \in \mathcal{G}_\beta$ such that $C_{G_\gamma}(v) = \gamma$ (proof in Appendix A.2).*

At the excluded boundaries, every affine downstream map with a nonzero baseline gap satisfies $C_G(v) = 0$ when $v^\top \delta_H = 0$ and $C_G(v) = 1$ when $v = \pm\delta_H/\|\delta_H\|_2$. The proposition establishes the insufficiency of the observational score law, not the necessity of post-erasure evaluations: when w and δ_H are known, Eq. (7) gives $C_G(v)$ directly.

4 OBERON: LEARNING DIRECTIONS FOR DETECTION AND INTERVENTION

Overview. The contrast between detection performance and reward-gap attenuation in Figure 1 motivates the joint objective of OBERON, which learns a single unit direction from matched hack–clean response pairs. The detection objective promotes positive paired score margins, while the intervention objective penalizes squared normalized reward residuals after erasing the same direction from every token state. We formulate a population risk combining these two criteria, derive a pairwise upper bound, and approximately minimize its empirical sequence-level instantiation to obtain $\widehat{v}$. At evaluation time, $\widehat{v}$ determines both the detector score $s_{\widehat{v}}$ and the erasure operator $Q_{\widehat{v}}$.

4.1 A PAIRWISE SURROGATE

Let π denote the fixed coupling of $P(H \mid Z = 1)$ and $P(H \mid Z = 0)$ induced by same-prompt matching, with $(H^+, H^-) \sim \pi$. For a unit direction v, define the paired detection margin $m_v := v^\top(H^+ - H^-)$ and the post-erasure reward residual $r_v := G(Q_v H^+) - G(Q_v H^-)$. The class marginals imply $\mathbb{E}_\pi[m_v] = D(v) = \mathcal{J}(s_v, I)$ and $\mathbb{E}_\pi[r_v] = \Delta_v = \mathcal{J}(G, Q_v)$. These first moments depend only on the marginals, whereas $\Pr_\pi(m_v \le 0)$ and $\mathbb{E}_\pi[r_v^2]$ also depend on the coupling. Same-prompt matching is therefore part of the method.

Definition 3 (Population risk). *For a fixed reward scale $s_g > 0$ and trade-off parameter $\alpha \ge 0$, define the population risk of a unit direction v as*

$$\mathcal{R}_\alpha(v) := \left(\frac{\Delta_v}{s_g}\right)^2 + \alpha \Pr_\pi(m_v \le 0). \tag{8}$$

The two terms measure the squared normalized mean reward gap after erasure and the paired misranking probability, respectively, with ties counted as errors.

Proposition 2 (Upper bound on the population risk). *Let $\ell(m) := \log(1 + e^{-m})/\log 2$ be the normalized logistic loss. Fix $s_g > 0$ and $\alpha \ge 0$. For every unit direction v satisfying $\mathbb{E}_\pi[r_v^2] < \infty$ and $\mathbb{E}_\pi[\ell(m_v)] < \infty$, the population risk admits the upper bound (proof in Appendix A.3):*

$$\mathcal{R}_\alpha(v) \le \mathbb{E}_\pi\left[\left(\frac{r_v}{s_g}\right)^2 + \alpha\,\ell(m_v)\right]. \tag{9}$$

We use the right-hand side of Eq. (9) as a population-level surrogate for direction learning. Its intervention term equals the squared normalized mean residual plus $\mathrm{Var}_\pi(r_v/s_g)$, penalizing variation across matched pairs and preventing cancellation between positive and negative residuals. The normalized logistic loss upper-bounds the paired misranking indicator while retaining sensitivity to the margin. This upper-bound relation does not imply that the surrogate and $\mathcal{R}_\alpha$ share a minimizer.

4.2 SEQUENCE-LEVEL INSTANTIATION

The preceding upper bound is formulated for vector-valued representations, whereas the selected model layer produces a token-state matrix $S \in \mathbb{R}^{T\times d}$. Let $M(S) \in \mathbb{R}^d$ denote its attention-mask mean and let $R(S)$ denote the frozen reward continuation with the input's attention mask held fixed. For each matched pair (S_i^+, S_i^-), write $m_i(v) := v^\top[M(S_i^+) - M(S_i^-)]$ for the detection margin and $r_i(v) := R(S_i^+ Q_v) - R(S_i^- Q_v)$ for the post-erasure reward residual, where $S_i^\pm Q_v = S_i^\pm - S_i^\pm v v^\top$ applies the erasure at every token position. For k matched pairs, let $g_i := R(S_i^+) - R(S_i^-)$ and set $s_g := \sqrt{k^{-1}\sum_{i=1}^k g_i^2}$. We assume $s_g > 0$ and hold it fixed with respect to v.

Definition 4 (Sequence-level OBERON objective). *For $\rho \in [0, 1)$, set $\alpha := \rho \log 2/(1-\rho)$. Replacing the expectation on the right-hand side of Equation* (9) *by the empirical mean and multiplying by the*

positive constant $1-\rho$ *gives the sequence-level objective to minimize over unit directions*

$$\widehat{\mathcal{L}}_\rho(v) := \frac{1}{k}\sum_{i=1}^{k}\left[(1-\rho)\left(\frac{r_i(v)}{s_g}\right)^2 + \rho\log(1+e^{-m_i(v)})\right]. \tag{10}$$

At $\rho = 1$, the objective extends continuously to the detection-only loss used in the ablation. We approximately minimize $\widehat{\mathcal{L}}_\rho(v)$ by applying Adam to u with $v = u/\|u\|_2$, while keeping the reward model frozen. Pooled detection states and unedited rewards are cached, whereas $r_i(v)$ is recomputed with Q_v applied at every token position. For evaluation, the fitted direction, denoted by $\widehat{v}$ after orientation, is assigned a nonnegative mean paired margin over the full direction-training partition, rather than only the k fitting pairs; its sign is retained when this mean is zero. Since $Q_{-v} = Q_v$, this convention leaves the intervention unchanged and also applies at $\rho = 0$.

5 Experiments

5.1 Experimental Setup

Data and reward models. We construct same-prompt pairs from RM-Bench (Liu et al., 2025b) to compare formatting (`fmt`), length (`len`), and quality–style (`cstyle`), and from SycophancyEval (Sharma et al., 2024) to compare incorrect agreement with correction (`syco`). The `fmt` and `len` pairs compare variants of chosen responses without independent quality re-annotation. The `syco` and `cstyle` pairs additionally vary correctness or chosen/rejected status, so attenuation is not inherently desirable. We evaluate all four conditions on five frozen reward models, GRM-Gemma2-2B, GRM-Llama3.1-8B, Skywork-Reward-8B, Skywork-Reward-V2-Llama-8B, and Skywork-Reward-V2-Qwen3-8B. Collateral quality is assessed on separate preference panels from Skywork-Reward-Preference-80K-v0.2 (Liu et al., 2024a), excluded from direction fitting.

Baselines. We compare difference of means (Arditi et al., 2024), the logistic probe (Alain & Bengio, 2016), and reward gradient, using their directions as linear readers and for the same rank-one orthogonal erasure as Oberon. LEACE (Belrose et al., 2023) uses its native erasure operator. Score penalty (Papadatos & Freedman, 2024) shares the logistic-probe reader but modifies output rewards rather than hidden states. Random directions and label-shuffled mean differences provide controls.

Evaluation metrics. We evaluate detection using AUROC on pre-intervention reader scores, with each reader oriented by the mean hack–clean score difference on the direction-training split. For n test pairs, let s_i^+ and s_i^- denote the oriented hack and clean scores, and let g_i and g_i' denote the hack–clean reward gaps before and after intervention. Unlike the theoretical signed ratio $C(v)$, we measure attenuation of absolute pairwise gaps using $C_{\text{pair}} := 100(1 - \sum_i |g_i'| / \sum_i |g_i|)$ for $\sum_i |g_i| > 0$, with negative values indicating amplification. Collateral quality is measured by Δ_{qual}, the pre-minus-post preference accuracy on the held-out quality panel in percentage points, with positive values indicating degradation. To assess whether detection and attenuation occur on the same pairs, we report the *pairwise detection-and-reduction rate* (PDR), the percentage of all test pairs with correct paired detection and a reduced absolute reward gap. Both criteria require strict inequalities, excluding score and gap ties from joint successes, giving

$$\text{PDR} := \frac{100}{n}\sum_{i=1}^{n}\mathbf{1}\{s_i^+ > s_i^-\}\mathbf{1}\{|g_i'| < |g_i|\}. \tag{11}$$

Evaluation protocol. We evaluate interventions on fixed response pairs without further policy optimization. Data are partitioned by source example into direction-training, validation, and test sets. Except for the random and label-shuffled controls, the main comparison uses three split seeds and three draws per split, each sampling $k = 40$ direction-training pairs without replacement. Splits and sampled pairs are shared across methods. Reader orientation is determined on the direction-training set, while the score-penalty coefficient is calibrated on the validation set. The main comparison uses decoder-layer output index 12 (zero-based), with $\rho = 0.8$ for Oberon. Readers use attention-mask mean-pooled hidden states, whereas representation-level interventions act on all token states at that

Table 1: **Detection, intervention, and collateral quality in reward-hacking evaluations.** Entries equally average five model-level estimates. AUROC measures detection before intervention, C_{pair} (%) quantifies absolute reward-gap attenuation, and PDR (%) measures the proportion of response pairs with both correct detection and absolute-gap reduction. Δ_{qual} reports the preference-accuracy drop in percentage points, with positive values indicating degradation. Bold denotes the highest PDR within each condition at the reported precision.

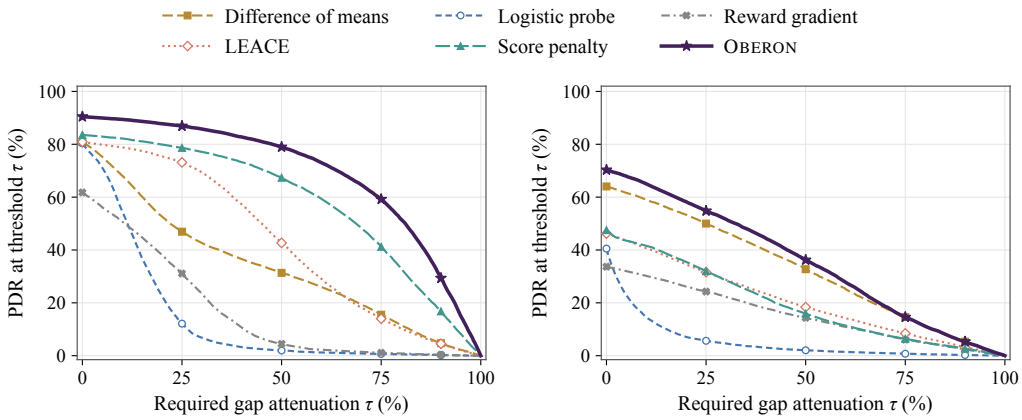
	Formatting (`fmt`)				Length (`len`)			
Method	AUROC ↑	C_{pair} ↑	PDR ↑	Δ_{qual} ↓	AUROC ↑	C_{pair} ↑	PDR ↑	Δ_{qual} ↓
Random direction	0.578	−0.2	26.8	−0.01	0.645	−0.2	28.7	−0.01
Label-shuffled mean difference	0.782	19.8	63.7	2.69	0.728	24.8	56.2	3.39
Detection baselines								
Difference of means	0.837	34.0	80.8	3.39	0.770	36.0	64.0	3.77
Logistic probe	0.974	10.4	80.3	0.01	0.892	−1.4	40.5	0.03
Intervention baselines								
Reward gradient	0.660	12.2	61.8	0.92	0.577	−3.3	33.7	0.88
LEACE	0.939	37.7	80.9	1.00	0.827	9.6	46.2	6.64
Score penalty	0.974	59.6	83.5	0.89	0.892	9.7	47.5	0.22
OBERON	0.945	72.7	**90.5**	0.31	0.842	40.9	**70.3**	0.91
	Sycophancy (`syco`)				Quality–style (`cstyle`)			
Method	AUROC ↑	C_{pair} ↑	PDR ↑	Δ_{qual} ↓	AUROC ↑	C_{pair} ↑	PDR ↑	Δ_{qual} ↓
Random direction	0.542	−0.6	32.0	−0.01	0.698	−0.2	30.6	0.02
Label-shuffled mean difference	0.831	−45.1	36.5	2.09	0.843	19.3	63.9	4.07
Detection baselines								
Difference of means	0.875	−32.1	61.7	2.13	0.895	20.7	62.8	3.73
Logistic probe	1.000	11.9	64.0	0.01	0.998	2.4	65.2	−0.02
Intervention baselines								
Reward gradient	0.601	25.0	64.4	0.19	0.691	11.8	43.6	0.65
LEACE	0.997	32.0	71.8	0.06	0.996	10.9	69.5	4.37
Score penalty	1.000	30.9	73.2	−0.03	0.998	24.7	66.4	0.61
OBERON	0.940	64.5	**82.6**	0.16	0.976	49.7	**78.9**	0.72

layer. Results are aggregated within each model and then averaged equally across models. Ablations use $k = 10$, with repetition schemes specified in the figure captions. Optimization details, control configurations, and metric-specific aggregation rules are provided in Appendix C.1.

5.2 MAIN RESULTS

Detection performance. To assess the discriminative ability of the jointly learned direction, we measure AUROC from reader scores on held-out responses before intervention and average the estimates equally across the five reward models. As shown in Table 1, OBERON attains AUROC values of 0.945 and 0.842 on `fmt` and `len`, respectively. For comparison, difference of means (Arditi et al., 2024) achieves 0.837 and 0.770, while the logistic probe (Alain & Bengio, 2016) achieves 0.974 and 0.892. The same ordering of average AUROC holds on `syco` and `cstyle`, where OBERON reaches 0.940 and 0.976, respectively. These results indicate that the jointly learned direction distinguishes the evaluated contrasts, although its detection performance alone does not establish its suitability for intervention. Detailed results for each reward model are provided in Appendix C.2.1.

Reward-gap attenuation. We compare the detection results with intervention effects measured by C_{pair}. Table 1 reports average attenuation of 72.7% and 40.9% for OBERON on `fmt` and `len`, respectively. For the logistic probe (Alain & Bengio, 2016), projecting out its learned direction from



(a) Formatting (`fmt`) (b) Length (`len`)

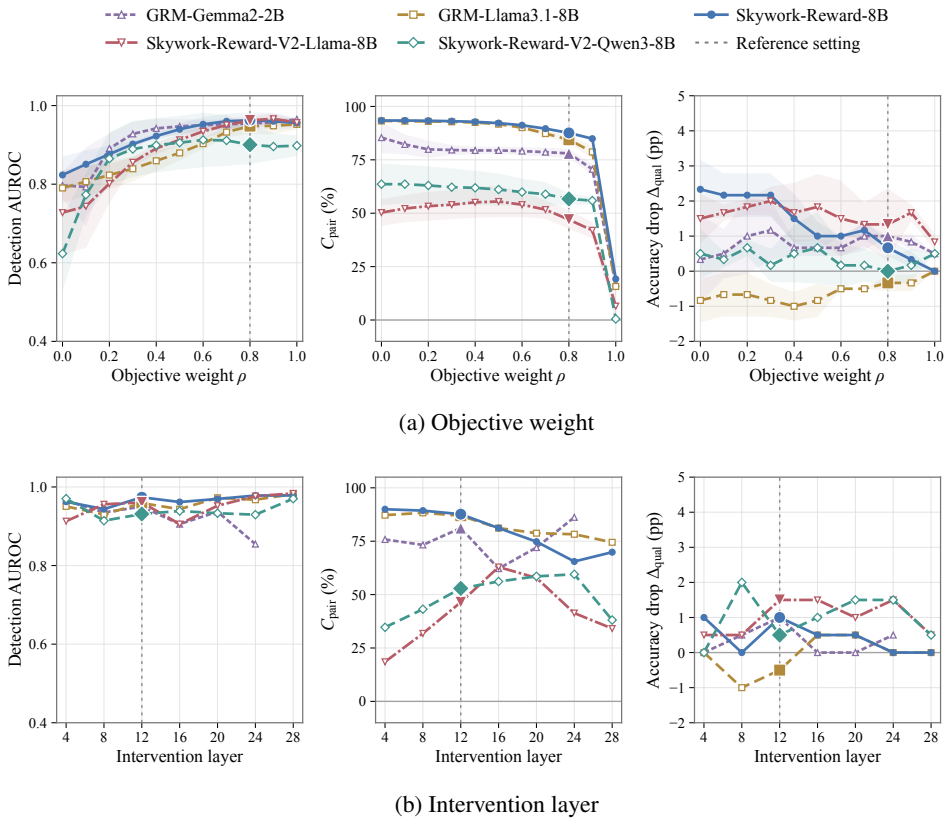
Figure 2: **PDR under stricter reward-gap attenuation requirements.** A test pair is successful at threshold τ if it is correctly detected and satisfies $|g_i'| < (1-\tau)|g_i|$, with all test pairs in the denominator and the original PDR recovered at $\tau = 0$. Curves average nine fit-level rates within each model and then equally across five models, keeping readers and interventions fixed across thresholds.

hidden representations yields attenuation of 10.4% and −1.4%. Thus, on both conditions, OBERON achieves greater average attenuation than logistic-probe projection despite its lower average AUROC. Score penalty (Papadatos & Freedman, 2024) uses the same reader as the logistic probe but modifies output rewards, achieving attenuation of 59.6% and 9.7%. This difference from projection shows that intervention effects also depend on the operator applied, even when the reader is fixed. On `syco` and `cstyle`, OBERON achieves average attenuation of 64.5% and 49.7%, respectively. These reductions do not necessarily indicate improved reward preferences because the conditions also vary correctness or preference quality. Appendix C.2.2 reports results for each model and scatter plots showing that aggregate attenuation can coexist with amplification on individual pairs.

Joint effectiveness and collateral quality. We evaluate the co-occurrence of correct detection and absolute reward-gap reduction using PDR, together with the associated preference-accuracy drop. Among the methods in Table 1, OBERON achieves the highest average PDR on all four conditions, reaching 90.5% on `fmt`, 70.3% on `len`, 82.6% on `syco`, and 78.9% on `cstyle`. Figure 2 evaluates whether this advantage persists under stricter attenuation requirements, with readers and interventions held fixed. On `fmt`, OBERON exceeds the five plotted baselines at all evaluated attenuation thresholds below 100%. On `len`, its advantage over difference of means narrows as the requirement increases, with the curves crossing at higher thresholds. The corresponding average preference-accuracy drops for OBERON are 0.31, 0.91, 0.16, and 0.72 percentage points across the four conditions. Although these averages are all below one percentage point, OBERON does not minimize collateral quality loss. On `len`, its PDR of 70.3% exceeds those of LEACE (Belrose et al., 2023) and score penalty (Papadatos & Freedman, 2024), at 46.2% and 47.5%, respectively. Its accuracy drop of 0.91 percentage points is smaller than LEACE's 6.64 but larger than score penalty's 0.22. These comparisons describe trade-offs at the evaluated settings rather than superiority at a matched quality cost. Threshold curves for all four conditions and per-model analyses are provided in Appendix C.2.3.

5.3 ABLATION STUDY

Objective weight. To examine the trade-off between detection and reward-gap attenuation, we vary ρ from 0 to 1 at layer 12. The endpoints correspond to removal-only and detection-only training, respectively. Figure 3(a) shows that, across five models on `fmt`, joint training at $\rho = 0.8$ achieves higher mean AUROC than removal-only training and higher mean C_{pair} than detection-only training. Increasing ρ beyond 0.8 yields little change in AUROC but reduces C_{pair}, most sharply at $\rho = 1$. These results support $\rho = 0.8$ as a detection-oriented compromise that retains substantial attenuation. Full `fmt` and `len` sweeps and model-specific comparisons are provided in Appendix C.2.5.



(a) Objective weight

(b) Intervention layer

Figure 3: **Sensitivity to objective weight and intervention layer.** Detection AUROC, reward-gap attenuation C_{pair}, and preference-accuracy drop (percentage points) on `fmt` across five reward models. (a) The weight sweep fixes layer 12, with curves showing means and shading indicating $\pm$ one standard deviation across three split seeds. (b) The layer sweep fixes $\rho = 0.8$, with a single fit per layer. Layer indices are zero-based, and layer 28 is unavailable for the 26-layer GRM-Gemma2-2B.

Intervention layer. To assess sensitivity to intervention location, we independently fit and apply a direction at each evaluated layer with $\rho = 0.8$, using one fit per model and layer. Figure 3(b) shows that later layers can weaken attenuation despite maintaining high AUROC, while changes in collateral quality remain model-dependent. On `fmt`, layer 12 retains high AUROC and substantial attenuation across the five models, with observed preference-accuracy drops of at most one percentage point. These observations support layer 12 as a common reference setting but do not establish it as a uniformly optimal intervention location. Full `fmt` and `len` results are provided in Appendix C.2.5.

6 Conclusion

In this work, we introduced OBERON, which jointly learns a single representation direction for bias detection and reward-gap attenuation in frozen reward models. To characterize the relationship between detection and erasure, we formulated an observation–intervention framework and established that detector scores alone do not generally identify signed erasure effects over a family of affine downstream maps. Within this framework, we derived a pairwise upper bound on a joint population risk, yielding an empirical objective that couples paired detection with post-erasure reward residuals. Experiments on fixed response pairs across five reward models show that OBERON achieves higher model-averaged pairwise detection-and-reduction rates than the evaluated baselines, with small average preference-accuracy drops. These findings suggest that direction learning should account for intervention effects rather than relying on detection performance alone. Future work could evaluate these interventions in best-of-N selection and policy optimization to assess whether reward-gap attenuation reduces reward exploitation while preserving response quality.

AI USE STATEMENT

Generative AI tools were used to review manuscript clarity and terminology and to check proofs, methodology, code, numerical results, figures, and interpretations. The authors considered the feedback and take responsibility for the final content and claims.

ETHICS STATEMENT

Reward hacking can increase reward scores without improving response quality and may reinforce misleading outputs. Our method aims to identify and reduce the influence of potentially exploitable reward-model signals. As our evaluation is limited to fixed response pairs, downstream use requires further assessment of response reliability and potential harms.

REPRODUCIBILITY STATEMENT

To support reproducibility, we describe the theoretical assumptions and method in Sections 3 and 4, with proofs and extensions in Appendices A and B. Data construction and experimental protocols are documented in Section 5.1 and Appendix C.1.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 136037–136083, 2024.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Lichang Chen, Chen Zhu, Jiuhai Chen, Davit Soselia, Tianyi Zhou, Tom Goldstein, Heng Huang, Mohammad Shoeybi, and Bryan Catanzaro. ODIN: Disentangled reward mitigates hacking in RLHF. In *International Conference on Machine Learning (ICML)*, volume 235, pp. 7935–7952, 2024.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. In *International Conference on Learning Representations (ICLR)*, volume 2024, pp. 50905–50931, 2024.

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

Daniel Fein, Max Lamparth, Violet Xiang, Mykel Kochenderfer, and Nick Haber. One bias after another: Mechanistic reward shaping and persistent biases in language reward models. In *International Conference on Machine Learning (ICML)*, 2026.

Cosimo Galeone, Anna Ettorre, Minsu Park, Giuseppe Ettorre, and Daniele Ligorio. Perfect detection, failed control: The geometry of knowing vs. steering in language models. *arXiv preprint arXiv:2606.24952*, 2026.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning (ICML)*, 2023.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019.

Nathan Lambert. Reinforcement learning from human feedback. *arXiv preprint arXiv:2504.12501*, 2025.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. Skywork-Reward: Bag of tricks for reward modeling in LLMs. *arXiv preprint arXiv:2410.18451*, 2024a.

Shuang Liu, Yuxuan Bo, Qiuyang Zhao, Caiyue Huang, Xiaorong Chen, Yanguang Liu, and Mengnan Du. HARVE: Hacking-aware reward-head vector editing for robust reward models. *arXiv preprint arXiv:2606.03131*, 2026.

Tianqi Liu, Wei Xiong, Jie Ren, Lichang Chen, Junru Wu, Rishabh Joshi, Yang Gao, Jiaming Shen, Zhen Qin, Tianhe Yu, Daniel Sohn, Anastasia Makarova, Jeremiah Zhe Liu, Yuan Liu, Bilal Piot, Abe Ittycheriah, Aviral Kumar, and Mohammad Saleh. RRM: Robust reward model training mitigates reward hacking. In *International Conference on Learning Representations (ICLR)*, pp. 62682–62700, 2025a.

Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. RM-Bench: Benchmarking reward models of language models with subtlety and style. In *International Conference on Learning Representations (ICLR)*, 2025b.

Zhihan Liu, Miao Lu, Shenao Zhang, Boyi Liu, Hongyi Guo, Yingxiang Yang, Jose Blanchet, and Zhaoran Wang. Provably mitigating overoptimization in rlhf: Your sft loss is implicitly an adversarial regularizer. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 138663–138697, 2024b.

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

Henry Papadatos and Rachel Freedman. Linear probe penalties reduce LLM sycophancy. In *NeurIPS Workshop on Socially Responsible Language Modelling Research*, 2024. arXiv:2412.00967.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7237–7256, 2020.

Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askell, Sam Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Kazutoshi Shinoda, Kosuke Nishida, and Kyosuke Nishida. Debiasing reward models via causally motivated inference-time intervention. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 22474–22489, 2026.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 9460–9471, 2022.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? even simple baselines outperform sparse autoencoders. In *International Conference on Machine Learning (ICML)*, 2025.

Yupei Yang, Lin Yang, Wanxi Deng, Lin Qu, Fan Feng, Biwei Huang, Shikui Tu, and Lei Xu. Factored causal representation learning for robust reward modeling in RLHF. In *International Conference on Machine Learning (ICML)*, 2026.

Xuanchang Zhang, Wei Xiong, Lichang Chen, Tianyi Zhou, Heng Huang, and Tong Zhang. From lists to emojis: How format bias affects model alignment. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 26940–26961, 2025.

A Proofs and Theoretical Extensions

A.1 Proof of Lemma 1

Lemma 1 (Factorization under an affine downstream map)**.** *Let $G(h) = c + w^\top h$, where $c \in \mathbb{R}$ and $w \in \mathbb{R}^d$. For any unit direction v, we have $\Delta_0 = w^\top \delta_H$ and $\Delta_0 - \Delta_v = D(v)(w^\top v)$. Whenever $\Delta_0 \neq 0$, the signed removal ratio factorizes as follows:*

$$C(v) = \frac{D(v)(w^\top v)}{\Delta_0} = \frac{(v^\top \delta_H)(w^\top v)}{w^\top \delta_H}. \tag{12}$$

Proof. Since the constant c cancels and $Q_v = I - vv^\top$,

$$\begin{aligned} \Delta_0 &= \mathcal{J}(G, I) = w^\top \delta_H, \\ \Delta_v &= \mathcal{J}(G, Q_v) = w^\top \delta_H - (w^\top v)(v^\top \delta_H), \\ \Delta_0 - \Delta_v &= (v^\top \delta_H)(w^\top v) = D(v)(w^\top v). \end{aligned}$$

Whenever $\Delta_0 \neq 0$,

$$C(v) = \frac{\Delta_0 - \Delta_v}{\Delta_0} = \frac{D(v)(w^\top v)}{\Delta_0} = \frac{(v^\top \delta_H)(w^\top v)}{w^\top \delta_H}. \qquad \square$$

A.2 Proof of Proposition 1

Proposition 1 (Non-identifiability of removal from observational scores)**.** *Fix the joint law P of (H, Z), a unit direction v satisfying $0 < |v^\top \delta_H| < \|\delta_H\|_2$, and a scalar $\beta \neq 0$. Let $\mathcal{G}_\beta$ consist of affine maps $G_{c,w}(h) = c + w^\top h$, where $c \in \mathbb{R}$ and $w \in \mathbb{R}^d$ satisfies $w^\top \delta_H = \beta$, with no restriction on $\|w\|_2$. The signed removal ratio $C_G(v)$ is not point-identifiable over $\mathcal{G}_\beta$ from the joint law of $(s_v(H), Z)$ and the baseline gap. All maps in $\mathcal{G}_\beta$ share this law and satisfy $\mathcal{J}(G, I) = \beta$, yet for every $\gamma \in \mathbb{R}$ there exists $G_\gamma \in \mathcal{G}_\beta$ such that $C_{G_\gamma}(v) = \gamma$.*

Proof. The construction perturbs w orthogonally to δ_H, preserving the baseline gap while changing $w^\top v$. Let

$$a := v^\top \delta_H, \qquad u := v - \frac{a}{\|\delta_H\|_2^2}\delta_H, \qquad w_0 := \frac{\beta}{\|\delta_H\|_2^2}\delta_H.$$

The assumptions imply $a \neq 0$ and $a^2 < \|\delta_H\|_2^2$. Using $\|v\|_2 = 1$, we obtain

$$\begin{aligned} u^\top \delta_H &= v^\top \delta_H - \frac{a}{\|\delta_H\|_2^2}\delta_H^\top \delta_H = 0, \\ u^\top v &= v^\top v - \frac{a}{\|\delta_H\|_2^2}\delta_H^\top v = 1 - \frac{a^2}{\|\delta_H\|_2^2} > 0, \\ w_0^\top \delta_H &= \frac{\beta}{\|\delta_H\|_2^2}\delta_H^\top \delta_H = \beta. \end{aligned}$$

For $w_t := w_0 + tu$, $t \in \mathbb{R}$,

$$\begin{aligned} w_t^\top \delta_H &= w_0^\top \delta_H + t\, u^\top \delta_H = \beta, \\ w_t^\top v &= w_0^\top v + t\, u^\top v. \end{aligned}$$

Thus, $G_{0,w_t} \in \mathcal{G}_\beta$ for every t, while $w_t^\top v$ ranges over $\mathbb{R}$. Every $G_{c,w} \in \mathcal{G}_\beta$ satisfies $\mathcal{J}(G_{c,w}, I) = w^\top \delta_H = \beta$, and the joint law of $(s_v(H), Z)$ is identical throughout $\mathcal{G}_\beta$ because P and v are fixed.

For any $\gamma \in \mathbb{R}$, define

$$t_\gamma := \frac{\gamma\beta/a - w_0^\top v}{u^\top v}, \qquad G_\gamma := G_{0,w_{t_\gamma}}.$$

These quantities are well defined because $a \neq 0$ and $u^\top v > 0$, and $G_\gamma \in \mathcal{G}_\beta$. Substitution and Eq. (7) give

$$w_{t_\gamma}^\top v = w_0^\top v + t_\gamma u^\top v = \frac{\gamma\beta}{a},$$
$$C_{G_\gamma}(v) = \frac{a(w_{t_\gamma}^\top v)}{\beta} = \gamma.$$

Since γ is arbitrary, the signed removal ratio can take any real value while the observational score law and baseline gap remain unchanged, establishing non-identifiability. □

A.3 Proof of Proposition 2

Proposition 2 (Upper bound on the population risk)**.** *Let $\ell(m) := \log(1 + e^{-m})/\log 2$ be the normalized logistic loss. Fix $s_g > 0$ and $\alpha \geq 0$. For every unit direction v satisfying $\mathbb{E}_\pi[r_v^2] < \infty$ and $\mathbb{E}_\pi[\ell(m_v)] < \infty$, the population risk admits the upper bound:*

$$\mathcal{R}_\alpha(v) \leq \mathbb{E}_\pi\left[\left(\frac{r_v}{s_g}\right)^2 + \alpha\,\ell(m_v)\right]. \tag{13}$$

Proof. Since ℓ is nonnegative and decreasing with $\ell(0) = 1$, it satisfies $\mathbf{1}\{m \leq 0\} \leq \ell(m)$ for every $m \in \mathbb{R}$. Using $\mathbb{E}_\pi[r_v] = \Delta_v$, $\Pr_\pi(m_v \leq 0) = \mathbb{E}_\pi[\mathbf{1}\{m_v \leq 0\}]$, and Jensen's inequality, we obtain

$$\begin{aligned}
\mathcal{R}_\alpha(v) &= \left(\frac{\Delta_v}{s_g}\right)^2 + \alpha \Pr_\pi(m_v \leq 0) \\
&= \left(\mathbb{E}_\pi\left[\frac{r_v}{s_g}\right]\right)^2 + \alpha\,\mathbb{E}_\pi[\mathbf{1}\{m_v \leq 0\}] \\
&\leq \mathbb{E}_\pi\left[\left(\frac{r_v}{s_g}\right)^2\right] + \alpha\,\mathbb{E}_\pi[\ell(m_v)] \\
&= \mathbb{E}_\pi\left[\left(\frac{r_v}{s_g}\right)^2 + \alpha\,\ell(m_v)\right].
\end{aligned}$$

□

B Beyond Affine Heads: A Path-Integral Representation

The encoding–use structure in Lemma 1 has a per-example analogue for differentiable nonlinear maps. For a unit v and $h \in \mathbb{R}^d$, define the erasure path $\gamma_v(t; h) := Q_v h + t\, vv^\top h$, $t \in [0, 1]$, connecting $\gamma_v(0; h) = Q_v h$ to $\gamma_v(1; h) = h$.

Proposition 3 (Path-integral representation)**.** *Fix a unit v and $h \in \mathbb{R}^d$, and suppose G is continuously differentiable on an open set containing the segment $\{\gamma_v(t; h) : t \in [0, 1]\}$. Then*

$$G(h) - G(Q_v h) = \Big[\int_0^1 v^\top \nabla G\big(\gamma_v(t; h)\big)\, dt\Big]\,(v^\top h). \tag{14}$$

If, in addition, the differentiability condition holds for $P(H \mid Z{=}z)$-almost every h and the right-hand side is integrable under both conditional laws, $z \in \{0, 1\}$, then

$$\Delta_0 - \Delta_v = \Delta_Z \mathbb{E}\Big[(v^\top H) \int_0^1 v^\top \nabla G\big(\gamma_v(t; H)\big)\, dt\Big],$$

where $\Delta_Z \mathbb{E}[\psi(H)] := \mathbb{E}[\psi(H) \mid Z{=}1] - \mathbb{E}[\psi(H) \mid Z{=}0]$.

Proof. Under the stated condition, $g_h(t) := G(\gamma_v(t; h))$ is continuously differentiable on $[0, 1]$ with $g_h'(t) = \nabla G(\gamma_v(t; h))^\top vv^\top h = \big[v^\top \nabla G(\gamma_v(t; h))\big](v^\top h)$ by the chain rule, since $\frac{d}{dt}\gamma_v(t; h) = vv^\top h$. The fundamental theorem of calculus gives $g_h(1) - g_h(0) = \int_0^1 g_h'(t)\, dt$, which is Eq. (14). Taking the class contrast $\Delta_Z \mathbb{E}[\cdot]$ of both sides gives the second display. □

Eq. (14) is a per-sample identity, not a first-order approximation: a nonzero per-sample erasure effect requires both a nonzero encoding $v^\top h$ and a nonzero *integrated* directional sensitivity along the erasure path. Two caveats keep the analogy with Lemma 1 honest. First, the product structure holds *per sample*: the population contrast is a class difference of expectations of products and in general does *not* factor into the mean encoding contrast times a single use coefficient. The affine case recovers Lemma 1 exactly because the sensitivity factor is constant there, $v^\top \nabla G \equiv v^\top w$, and pulls out of the expectation. Second, the identity concerns the output difference of the frozen head under representation erasure; we attach no structural-causal semantics to it.

C Additional Experiments

C.1 Experimental Setup

Data and pair construction. We retain 658 chat and mathematics examples from RM-Bench (Liu et al., 2025b) with the response variants required for the paired comparisons. Formatting (`fmt`) pairs a detailed Markdown chosen response with its detailed plain text counterpart, while length (`len`) pairs the detailed plain text chosen response with its concise counterpart. Quality–style (`cstyle`) pairs the detailed Markdown rejected response with the concise chosen response from the same example. For sycophancy (`syco`), we use the first 543 records from SycophancyEval's `are_you_sure` subset (Sharma et al., 2024), all containing multiple-choice questions. We append an incorrect proposed answer to each question and pair a templated agreement response with a corrective response. The first and second responses in each pair are operationally labeled hack and clean, respectively. These labels define the contrast rather than an independently verified quality ordering. Specifically, `fmt` and `len` compare variants of chosen responses without additional quality annotation, whereas the paired responses differ in correctness on `syco` and in chosen/rejected status on `cstyle`. Collateral quality is evaluated on separate chosen–rejected panels from Skywork-Reward-Preference-80K-v0.2 (Liu et al., 2024a), excluded from direction fitting. For learned methods, the main evaluations use a fixed quality panel of 1,000 pairs, except for `cstyle` on GRM-Gemma2-2B, Skywork-Reward-V2-Llama-8B, and Skywork-Reward-V2-Qwen3-8B, which use panels of 300 pairs. The ablation sweeps use quality panels of 200 pairs.

Reward models. We evaluate GRM-Gemma2-2B, GRM-Llama3.1-8B, Skywork-Reward-8B, Skywork-Reward-V2-Llama-8B, and Skywork-Reward-V2-Qwen3-8B, covering the Gemma, Llama, and Qwen families. All five models are evaluated on all four conditions. All reward model parameters remain frozen, with the models in evaluation mode throughout direction fitting and intervention. Layer indices start at zero and refer to decoder block outputs, excluding the embedding output. GRM-Gemma2-2B has 26 decoder layers, the three Llama-based models have 32, and Skywork-Reward-V2-Qwen3-8B has 36. The intervention index of 12 used in the main experiments does not correspond to the same relative depth across models.

Baseline implementations. Difference of means (Arditi et al., 2024) estimates a direction from the average paired difference in hidden states. The logistic probe (Alain & Bengio, 2016) is fitted on standardized hidden states. Let w_{std} denote its fitted coefficient vector and σ_j the scaling factor for feature j. We recover the coefficients in the original coordinates as $a_j = (w_{\mathrm{std}})_j/\sigma_j$ and, for $a \neq 0$, use $v = a/\|a\|_2$ for both reading and erasure. Reward gradient averages gradients of hack-response rewards with respect to hidden states at the intervention layer, first over non-padding tokens and then over fitting examples. These directions define linear readers and use the same orthogonal erasure operator Q_v as OBERON. LEACE (Belrose et al., 2023) uses its fitted erasure operator, whose linear part is represented as $I - AB$. The factors A and B are stored as `proj_left` and `proj_right`, respectively. For detection, we normalize the first row of B to unit length and orient the resulting reader on the direction-training partition. Detection is measured before erasure and does not assess residual concept recoverability. For LEACE's `syco` and `cstyle` PDR evaluations, readers are reconstructed from the saved erasure artifacts. Score penalty adapts probe-based reward penalization (Papadatos & Freedman, 2024), reusing the logistic-probe reader and subtracting a validation-calibrated multiple of its score from the output reward without editing hidden states. Random directions provide a projection control. Label-shuffled mean differences are estimated by randomly swapping hack and clean labels within direction-training pairs. The endpoint $\rho = 0$ of the OBERON objective provides the removal-only ablation, while $\rho = 1$ provides detection-only training.

Table 2: **Detection, intervention, and collateral quality for GRM-Gemma2-2B.** C_{pair} and PDR are percentages. Δ_{qual} denotes the drop in preference accuracy in percentage points. Bold marks the highest displayed PDR, including controls. Shading identifies OBERON. Higher attenuation or PDR on `syco` and `cstyle` does not necessarily indicate a better outcome.

	Formatting (`fmt`)				Length (`len`)			
Method	AUROC ↑	C_{pair} ↑	PDR ↑	Δ_{qual} ↓	AUROC ↑	C_{pair} ↑	PDR ↑	Δ_{qual} ↓
Random direction	0.555	−0.3	25.6	−0.04	0.572	−0.6	18.7	−0.04
Label-shuffled mean difference	0.828	1.2	52.7	0.30	0.701	−4.7	40.0	1.35
Detection baselines								
Difference of means	0.889	12.5	87.1	0.53	0.756	14.1	**56.3**	1.73
Logistic probe	0.980	18.8	**96.9**	−0.03	0.891	−2.1	39.9	−0.07
Intervention baselines								
Reward gradient	0.657	22.9	87.3	0.97	0.518	−2.5	34.4	1.00
LEACE	0.950	55.3	96.8	0.00	0.838	3.4	44.1	0.67
Score penalty	0.980	70.0	92.3	0.87	0.891	6.5	44.5	0.20
OBERON	0.959	77.9	95.4	0.50	0.870	10.4	54.7	0.23

	Sycophancy (`syco`)				Quality–style (`cstyle`)			
Method	AUROC ↑	C_{pair} ↑	PDR ↑	Δ_{qual} ↓	AUROC ↑	C_{pair} ↑	PDR ↑	Δ_{qual} ↓
Random direction	0.580	−1.2	29.4	−0.04	0.580	−0.3	23.8	−0.08
Label-shuffled mean difference	0.923	−63.4	38.1	0.47	0.773	−17.0	31.2	1.10
Detection baselines								
Difference of means	0.958	−67.7	37.9	0.60	0.899	−9.0	30.2	1.89
Logistic probe	1.000	5.2	52.1	−0.07	0.997	4.1	74.5	0.00
Intervention baselines								
Reward gradient	0.577	39.8	**69.6**	1.23	0.629	5.4	50.0	0.00
LEACE	0.998	1.3	49.2	0.23	0.994	15.5	**85.4**	0.44
Score penalty	1.000	16.9	62.3	−0.03	0.997	41.8	73.5	0.89
OBERON	1.000	40.0	68.5	0.07	0.993	36.8	81.8	1.00

Training and evaluation protocol. For each of three split seeds, the RM-Bench examples are partitioned into 341 direction-training, 86 validation, and 231 test examples. The corresponding counts for `syco` are 281, 71, and 191. All variants of a source example remain in the same partition, and methods and conditions derived from the same source share splits. The main evaluations of learned methods use three draws per split, each sampling $k = 40$ pairs without replacement from the direction-training partition. The sampled pairs are shared across methods within each draw. Reader orientation is determined from scores on the full direction-training partition, without using test data. At decoder block output index 12, detection uses attention-mask mean pooling, while representation erasure is applied to every token state before the remaining reward computation. We parameterize the OBERON direction as $v = u/\|u\|_2$ and optimize u for 40 full-batch Adam steps with learning rate 0.05, using $\rho = 0.8$. The RMS scale s_g is computed from unedited reward gaps on the fitting pairs and held fixed during optimization. Detection and gap metrics for the random and label-shuffled controls use five replicates per control at split seed 1, with bfloat16 computation and at most 96 test pairs. Their quality drops come from separate run-level summaries rather than these reduced evaluations. The objective-weight and layer ablations use $k = 10$ pairs and the same 40-step optimization budget. The weight sweep evaluates $\rho \in \{0, 0.1, \ldots, 1\}$ at layer 12 using three split seeds and one draw per split. The layer sweep fixes $\rho = 0.8$ and independently fits the direction at layers $\{4, 8, 12, 16, 20, 24, 28\}$ using split seed 1 and draw 0. Layer 28 is excluded for GRM-Gemma2-2B because it is out of range. These sweeps assess sensitivity under the $k = 10$ configuration rather than directly reproduce the main results obtained with $k = 40$.

Table 3: **Detection, intervention, and collateral quality for GRM-Llama3.1-8B.** C_{pair} and PDR are percentages. Δ_{qual} denotes the drop in preference accuracy in percentage points. Bold marks the highest displayed PDR, including controls. Shading identifies OBERON. Higher attenuation or PDR on `syco` and `cstyle` does not necessarily indicate a better outcome.

	Formatting (`fmt`)				Length (`len`)			
Method	AUROC ↑	C_{pair} ↑	PDR ↑	Δ_{qual} ↓	AUROC ↑	C_{pair} ↑	PDR ↑	Δ_{qual} ↓
Random direction	0.576	−0.4	20.0	−0.02	0.659	0.0	31.0	−0.02
Label-shuffled mean difference	0.813	7.1	58.1	−0.01	0.742	14.6	58.5	1.03
Detection baselines								
Difference of means	0.873	10.9	66.2	0.37	0.792	23.4	61.4	1.30
Logistic probe	0.965	14.0	**96.6**	−0.07	0.893	0.2	46.8	0.03
Intervention baselines								
Reward gradient	0.700	−8.7	35.3	0.57	0.545	−5.0	39.2	0.47
LEACE	0.937	60.6	96.2	−0.27	0.825	11.5	53.2	0.80
Score penalty	0.965	66.1	84.0	1.83	0.893	0.8	43.0	0.10
OBERON	0.958	85.7	96.2	−0.30	0.833	56.4	**80.5**	0.63

	Sycophancy (`syco`)				Quality–style (`cstyle`)			
Method	AUROC ↑	C_{pair} ↑	PDR ↑	Δ_{qual} ↓	AUROC ↑	C_{pair} ↑	PDR ↑	Δ_{qual} ↓
Random direction	0.539	−0.5	28.8	−0.02	0.728	−0.4	24.2	−0.02
Label-shuffled mean difference	0.941	1.2	45.0	0.03	0.861	6.3	56.2	0.53
Detection baselines								
Difference of means	0.977	25.9	**94.0**	0.07	0.916	8.4	60.3	0.50
Logistic probe	1.000	10.8	74.6	−0.10	0.998	3.4	76.1	0.00
Intervention baselines								
Reward gradient	0.584	25.8	82.3	−0.13	0.673	−12.0	25.6	0.57
LEACE	0.994	52.4	91.8	0.00	0.996	26.7	**84.9**	−0.13
Score penalty	1.000	52.1	84.4	−0.13	0.998	35.3	71.8	1.30
OBERON	0.970	74.0	89.5	−0.30	0.968	67.6	83.7	0.10

Statistical reporting. PDR requires strictly positive paired detection margins and a strict reduction in the absolute reward gap, with all test pairs in the denominator. For the main evaluations of learned methods, AUROC, C_{pair}, and preference accuracy drops are summarized by the median over three draws within each split and then averaged across three splits. PDR instead averages the nine fit-level rates. Table 1 averages the five reported model-level estimates equally for each condition, including `cstyle`. The table uses rounded model summaries, whereas the PDR threshold curves average unrounded fit-level rates within each model and then equally across models. PDR values at $\tau = 0$ can therefore differ slightly from the corresponding table entries due to rounding. In the weight ablation, curves show means across three split seeds, and shaded bands indicate one descriptive standard deviation rather than confidence intervals. For this sweep, AUROC and preference accuracy drop statistics are recomputed from run records, whereas C_{pair} means and standard deviations are taken from the supplied summaries. The layer ablation uses one fit for each combination of model, condition, and layer, without estimates of variability across seeds. Repeated draws and splits may share test examples and are not treated as independent response samples. Quality drops are reported in percentage points, with negative values indicating observed accuracy improvements. Bold values identify the highest displayed PDR within each condition, not statistical significance.

C.2 EXPERIMENTAL RESULTS

We examine detection performance across the five reward models in Section C.2.1 and analyze attenuation of reward gaps and intervention effects on individual response pairs in Section C.2.2.

Table 4: **Detection, intervention, and collateral quality for Skywork-Reward-8B.** C_{pair} and PDR are percentages. Δ_{qual} denotes the drop in preference accuracy in percentage points. Bold marks the highest displayed PDR, including controls. Shading identifies OBERON. Higher attenuation or PDR on `syco` and `cstyle` does not necessarily indicate a better outcome.

	Formatting (`fmt`)				Length (`len`)			
Method	AUROC ↑	C_{pair} ↑	PDR ↑	Δ_{qual} ↓	AUROC ↑	C_{pair} ↑	PDR ↑	Δ_{qual} ↓
Random direction	0.578	−0.1	34.4	0.00	0.655	−0.3	31.5	0.00
Label-shuffled mean difference	0.798	27.3	79.4	0.18	0.738	39.4	72.7	1.78
Detection baselines								
Difference of means	0.893	71.8	96.1	1.07	0.790	56.9	78.7	3.83
Logistic probe	0.973	13.0	**96.5**	0.00	0.899	−2.8	35.5	−0.03
Intervention baselines								
Reward gradient	0.725	26.8	85.0	0.50	0.516	11.9	39.1	0.30
LEACE	0.940	49.4	95.8	0.13	0.812	−1.1	46.6	0.63
Score penalty	0.973	71.2	87.4	2.07	0.899	10.2	48.7	0.17
OBERON	0.967	86.5	95.2	0.10	0.831	57.5	**80.4**	0.80

	Sycophancy (`syco`)				Quality–style (`cstyle`)			
Method	AUROC ↑	C_{pair} ↑	PDR ↑	Δ_{qual} ↓	AUROC ↑	C_{pair} ↑	PDR ↑	Δ_{qual} ↓
Random direction	0.531	0.6	30.4	0.00	0.722	−0.1	32.9	0.00
Label-shuffled mean difference	0.913	−3.2	40.8	0.09	0.868	46.5	88.1	1.18
Detection baselines								
Difference of means	0.968	20.9	84.6	0.10	0.918	48.2	**91.3**	0.83
Logistic probe	1.000	28.2	80.4	0.00	0.998	2.5	69.1	0.00
Intervention baselines								
Reward gradient	0.626	60.6	**88.9**	0.27	0.747	23.7	64.4	0.37
LEACE	0.996	44.9	78.4	0.03	0.996	25.0	84.4	−0.03
Score penalty	1.000	40.0	77.5	0.00	0.998	37.2	73.7	1.07
OBERON	0.932	74.2	86.2	0.20	0.980	72.2	87.9	0.27

Section C.2.3 evaluates PDR under increasing attenuation thresholds and examines the relationship between original PDR and the drop in preference accuracy. We assess attenuation relative to changes in the scale of preference margins on a separate quality panel in Section C.2.4. Finally, Section C.2.5 examines sensitivity to the objective weight and intervention layer.

C.2.1 DETECTION PERFORMANCE

We compare OBERON with the two detection baselines, difference of means and the logistic probe, to examine whether the AUROC ordering in Table 1 holds for each reward model. On `fmt`, `len`, and `cstyle`, OBERON achieves higher AUROC than difference of means and lower AUROC than the logistic probe for all five models. Its AUROC ranges from 0.880 to 0.967 on `fmt`, from 0.830 to 0.870 on `len`, and from 0.963 to 0.993 on `cstyle`. The gap to the logistic probe nevertheless varies across models. On `fmt`, the AUROC gap ranges from 0.006 to 0.021 on four of the five models, but reaches 0.095 on Skywork-Reward-V2-Qwen3-8B. For this model, Table 6 reports an AUROC of 0.880 for OBERON, compared with 0.975 for the logistic probe and 0.668 for difference of means.

The comparison with difference of means is less uniform on `syco`. On GRM-Gemma2-2B, OBERON and the logistic probe both achieve 1.000 at the reported precision, compared with 0.958 for difference of means, as reported in Table 2. OBERON also exceeds difference of means on Skywork-Reward-V2-Llama-8B and Skywork-Reward-V2-Qwen3-8B, with respective comparisons of 0.962 versus 0.938 and 0.837 versus 0.535 in Tables 5 and 6. However, Tables 3 and 4 show the reverse ordering on GRM-Llama3.1-8B and Skywork-Reward-8B. On these two models, OBERON achieves 0.970

Table 5: **Detection, intervention, and collateral quality for Skywork-Reward-V2-Llama-8B.** C_{pair} and PDR are percentages. Δ_{qual} denotes the drop in preference accuracy in percentage points. Bold marks the highest displayed PDR, including controls. Shading identifies OBERON. Higher attenuation or PDR on `syco` and `cstyle` does not necessarily indicate a better outcome.

	Formatting (`fmt`)				Length (`len`)			
Method	AUROC ↑	C_{pair} ↑	PDR ↑	Δ_{qual} ↓	AUROC ↑	C_{pair} ↑	PDR ↑	Δ_{qual} ↓
Random direction	0.581	−0.5	19.2	0.03	0.670	0.0	25.0	0.03
Label-shuffled mean difference	0.808	4.3	43.5	0.40	0.733	18.1	57.5	0.91
Detection baselines								
Difference of means	0.861	16.8	70.9	0.90	0.770	29.4	65.4	0.13
Logistic probe	0.978	8.2	**82.7**	0.17	0.893	−1.6	38.2	0.13
Intervention baselines								
Reward gradient	0.612	27.4	68.2	2.10	0.576	19.6	33.3	1.77
LEACE	0.936	27.9	74.0	0.23	0.809	15.2	48.6	0.97
Score penalty	0.978	35.7	70.4	0.17	0.893	15.3	54.6	0.37
OBERON	0.963	54.0	76.6	1.33	0.845	48.3	**75.1**	1.20
	Sycophancy (`syco`)				Quality–style (`cstyle`)			
Method	AUROC ↑	C_{pair} ↑	PDR ↑	Δ_{qual} ↓	AUROC ↑	C_{pair} ↑	PDR ↑	Δ_{qual} ↓
Random direction	0.531	−0.2	37.1	0.03	0.728	0.0	32.3	0.03
Label-shuffled mean difference	0.847	0.6	50.4	−0.03	0.869	11.1	64.0	0.93
Detection baselines								
Difference of means	0.938	19.6	80.5	−0.20	0.894	6.3	59.5	−0.44
Logistic probe	1.000	16.4	77.3	0.27	0.998	2.6	63.1	−0.44
Intervention baselines								
Reward gradient	0.710	34.7	76.4	0.30	0.617	47.6	46.9	1.22
LEACE	0.995	29.9	71.3	0.33	0.996	5.3	55.1	−1.00
Score penalty	1.000	21.9	70.2	0.20	0.998	7.8	59.5	0.00
OBERON	0.962	72.3	**86.3**	0.80	0.977	47.2	**64.3**	0.44

and 0.932, respectively, compared with 0.977 and 0.968 for difference of means. Across all four conditions, OBERON therefore achieves higher AUROC than difference of means in 18 of the 20 combinations of model and condition. The logistic probe achieves higher AUROC than OBERON in 19 combinations and ties in one at the reported precision.

C.2.2 REWARD-GAP ATTENUATION

On both `fmt` and `len`, OBERON achieves greater C_{pair} than orthogonal erasure using the logistic probe direction, reward gradient intervention, LEACE, and score penalty for all five reward models. Although the logistic probe achieves higher AUROC, OBERON produces greater attenuation for every model on both conditions, showing a reversal in rankings. Attenuation varies across models, ranging from 54.0% to 86.5% on `fmt` and from 10.4% to 57.5% on `len`. Difference of means yields greater attenuation than OBERON on `len` for GRM-Gemma2-2B and Skywork-Reward-V2-Qwen3-8B, with respective comparisons of 14.1% versus 10.4% and 56.2% versus 31.9%.

Figures 4 through 8 compare absolute reward gaps before and after intervention across all five models. Each figure uses split seed 1 and draw 0, retaining all 231 test pairs per condition with shared normalization and axis limits across methods. These figures show individual fits rather than the aggregated estimates reported in the tables. Positive aggregate attenuation can coexist with amplification on individual pairs. For GRM-Gemma2-2B on `len`, OBERON reduces the absolute gap on 110 pairs, increases it on 108, and leaves 13 unchanged, while achieving $C_{\text{pair}} = 10.7\%$. For Skywork-Reward-V2-Qwen3-8B under the same condition, the corresponding counts are 145, 73, and

Table 6: **Detection, intervention, and collateral quality for Skywork-Reward-V2-Qwen3-8B.** C_{pair} and PDR are percentages. Δ_{qual} denotes the drop in preference accuracy in percentage points. Bold marks the highest displayed PDR, including controls. Shading identifies OBERON. Higher attenuation or PDR on `syco` and `cstyle` does not necessarily indicate a better outcome.

	Formatting (`fmt`)				Length (`len`)			
Method	AUROC ↑	C_{pair} ↑	PDR ↑	Δ_{qual} ↓	AUROC ↑	C_{pair} ↑	PDR ↑	Δ_{qual} ↓
Random direction	0.600	0.3	34.6	0.00	0.669	−0.1	37.3	0.00
Label-shuffled mean difference	0.665	58.9	84.8	12.58	0.728	56.7	52.5	11.89
Detection baselines								
Difference of means	0.668	57.9	83.6	14.07	0.740	56.2	58.4	11.87
Logistic probe	0.975	−1.9	29.0	0.00	0.884	−0.7	42.1	0.10
Intervention baselines								
Reward gradient	0.607	−7.2	33.0	0.47	0.730	−40.7	22.3	0.87
LEACE	0.931	−4.8	41.7	4.90	0.851	19.1	38.5	30.13
Score penalty	0.975	54.8	83.6	−0.50	0.884	15.5	46.7	0.27
OBERON	0.880	59.6	**88.9**	−0.10	0.830	31.9	**60.9**	1.70
	Sycophancy (`syco`)				Quality–style (`cstyle`)			
Method	AUROC ↑	C_{pair} ↑	PDR ↑	Δ_{qual} ↓	AUROC ↑	C_{pair} ↑	PDR ↑	Δ_{qual} ↓
Random direction	0.529	−1.9	34.2	0.00	0.730	0.0	39.6	0.15
Label-shuffled mean difference	0.530	−160.6	8.1	9.88	0.843	49.4	**80.0**	16.60
Detection baselines								
Difference of means	0.535	−159.1	11.7	10.10	0.847	49.5	72.8	15.89
Logistic probe	1.000	−1.2	35.8	−0.03	0.998	−0.8	43.0	0.33
Intervention baselines								
Reward gradient	0.509	−35.7	4.9	−0.70	0.791	−5.5	31.2	1.11
LEACE	1.000	31.3	68.3	−0.27	0.996	−18.0	37.9	22.56
Score penalty	1.000	23.5	71.7	−0.17	0.998	1.2	53.7	−0.22
OBERON	0.837	62.0	**82.4**	0.03	0.963	24.5	76.6	1.78

13, with $C_{\text{pair}} = 30.9\%$. Across methods, removal-only achieves greater aggregate attenuation than OBERON in all ten plotted combinations of model and condition. These comparisons characterize attenuation, while joint effectiveness and collateral quality require separate evaluation. We further examine attenuation relative to changes in the scale of preference margins in Appendix C.2.4.

C.2.3 JOINT EFFECTIVENESS AND COLLATERAL QUALITY

The PDR comparison with the logistic probe varies across models. On `fmt`, OBERON exceeds the logistic probe only on Skywork-Reward-V2-Qwen3-8B, with 88.9% versus 29.0% in Table 6, while trailing by 0.4 to 6.1 percentage points on the other four models. Its average PDR advantage over the logistic probe on `fmt` is therefore driven by the large gain on this model, rather than consistent gains across all five. On `len`, `syco`, and `cstyle`, OBERON achieves higher PDR than the logistic probe on every model, but does not achieve the highest PDR in every setting. For example, Table 2 shows that difference of means attains 56.3% PDR on `len`, compared with 54.7% for OBERON, while LEACE attains 85.4% on `cstyle`, compared with 81.8%. Figure 9 extends the threshold analysis to all four conditions, showing how average joint success changes as the required attenuation increases. Evaluating this criterion separately for each model shows that, at $\tau = 0.5$, OBERON achieves higher PDR than all five baselines on every model on `fmt`, with values ranging from 60.8% to 91.3%. On `len`, it exceeds difference of means on three of the five models at this threshold. Thus, method rankings under the original PDR criterion can differ from those requiring substantial gap reduction.

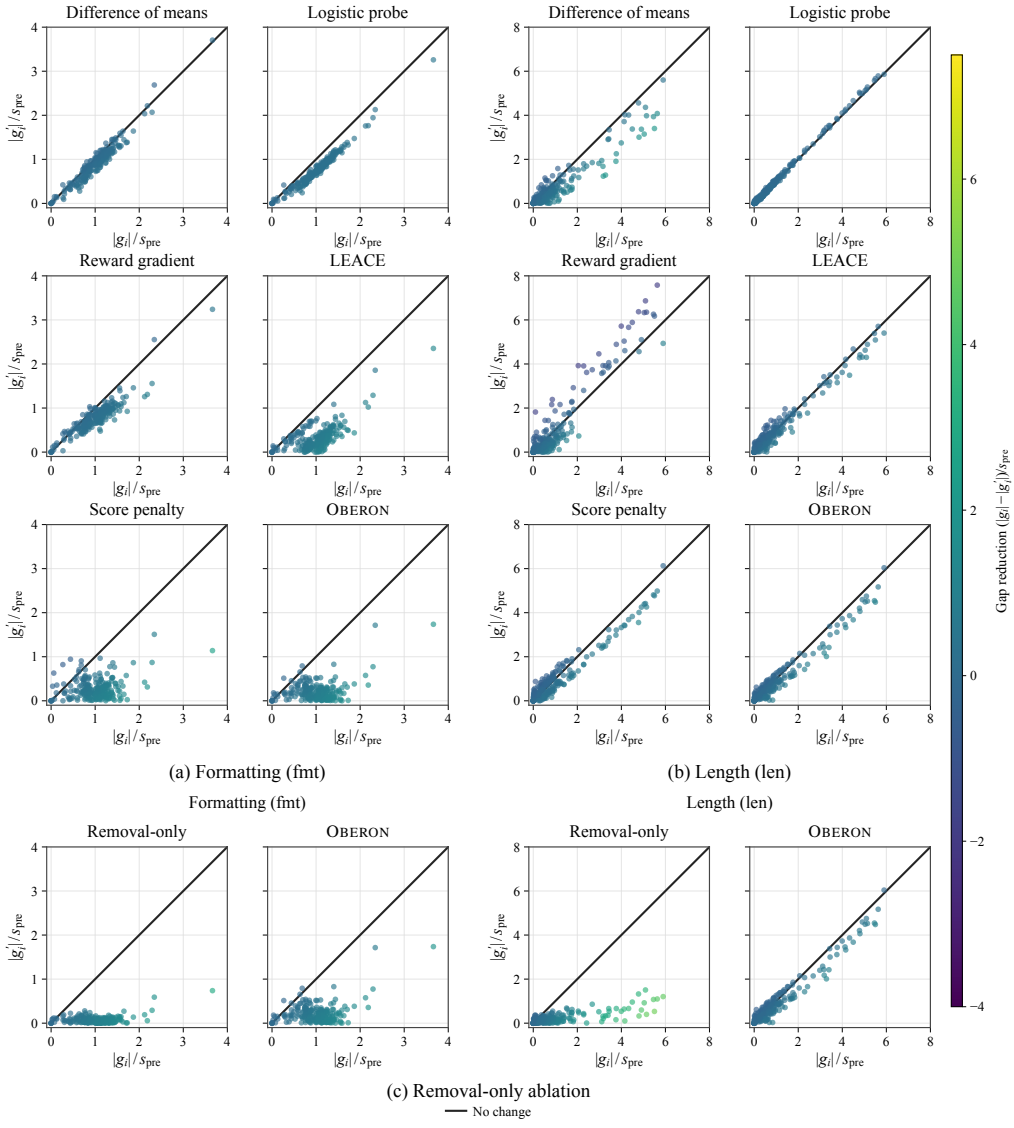


Figure 4: **Intervention effects on response pairs for GRM-Gemma2-2B.** Panels (a) and (b) compare OBERON with five baselines on `fmt` and `len`, respectively, while panel (c) compares OBERON with a variant trained using only the removal objective. The horizontal and vertical axes show absolute reward gaps before and after intervention, normalized by the mean absolute baseline gap s_{pre}. Color encodes $(|g_i| - |g_i'|)/s_{\text{pre}}$, with points below the diagonal indicating attenuation.

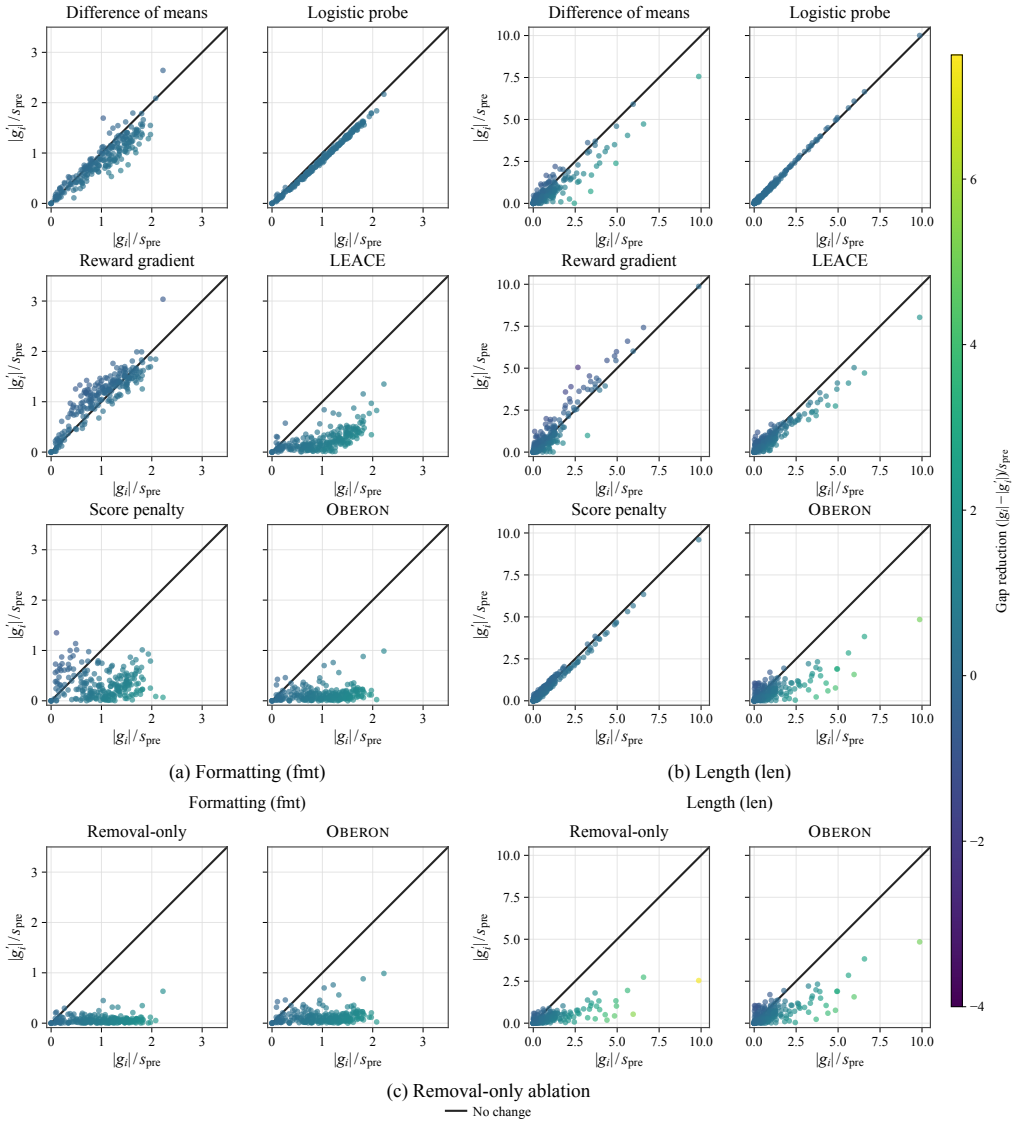


Figure 5: **Intervention effects on response pairs for GRM-Llama3.1-8B.** Panels (a) and (b) compare OBERON with five baselines on `fmt` and `len`, respectively, while panel (c) compares OBERON with a variant trained using only the removal objective. The horizontal and vertical axes show absolute reward gaps before and after intervention, normalized by the mean absolute baseline gap s_{pre}. Color encodes $(|g_i| - |g_i'|)/s_{\text{pre}}$, with points below the diagonal indicating attenuation.

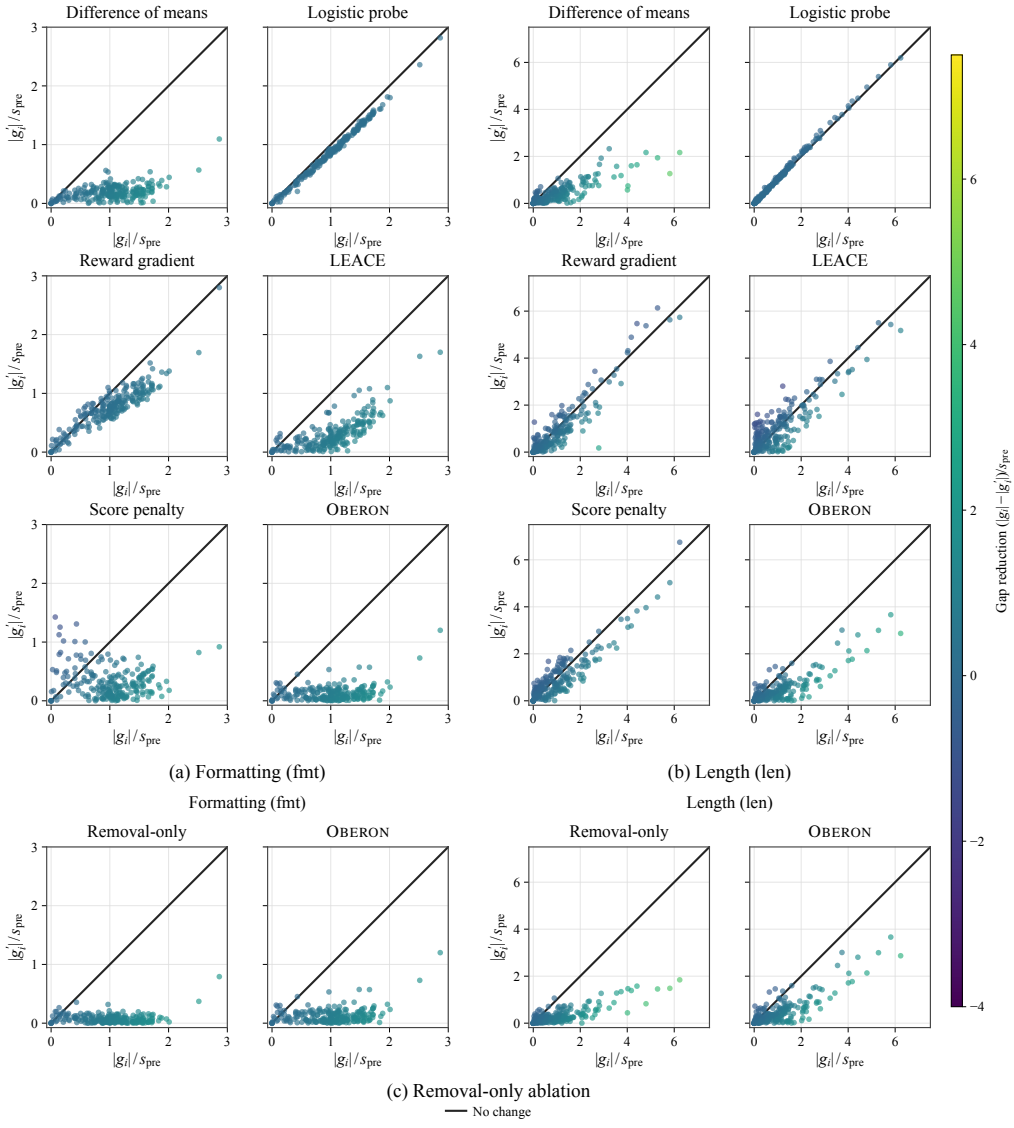


Figure 6: **Intervention effects on response pairs for Skywork-Reward-8B.** Panels (a) and (b) compare OBERON with five baselines on `fmt` and `len`, respectively, while panel (c) compares OBERON with a variant trained using only the removal objective. The horizontal and vertical axes show absolute reward gaps before and after intervention, normalized by the mean absolute baseline gap s_{pre}. Color encodes $(|g_i| - |g_i'|)/s_{\text{pre}}$, with points below the diagonal indicating attenuation.

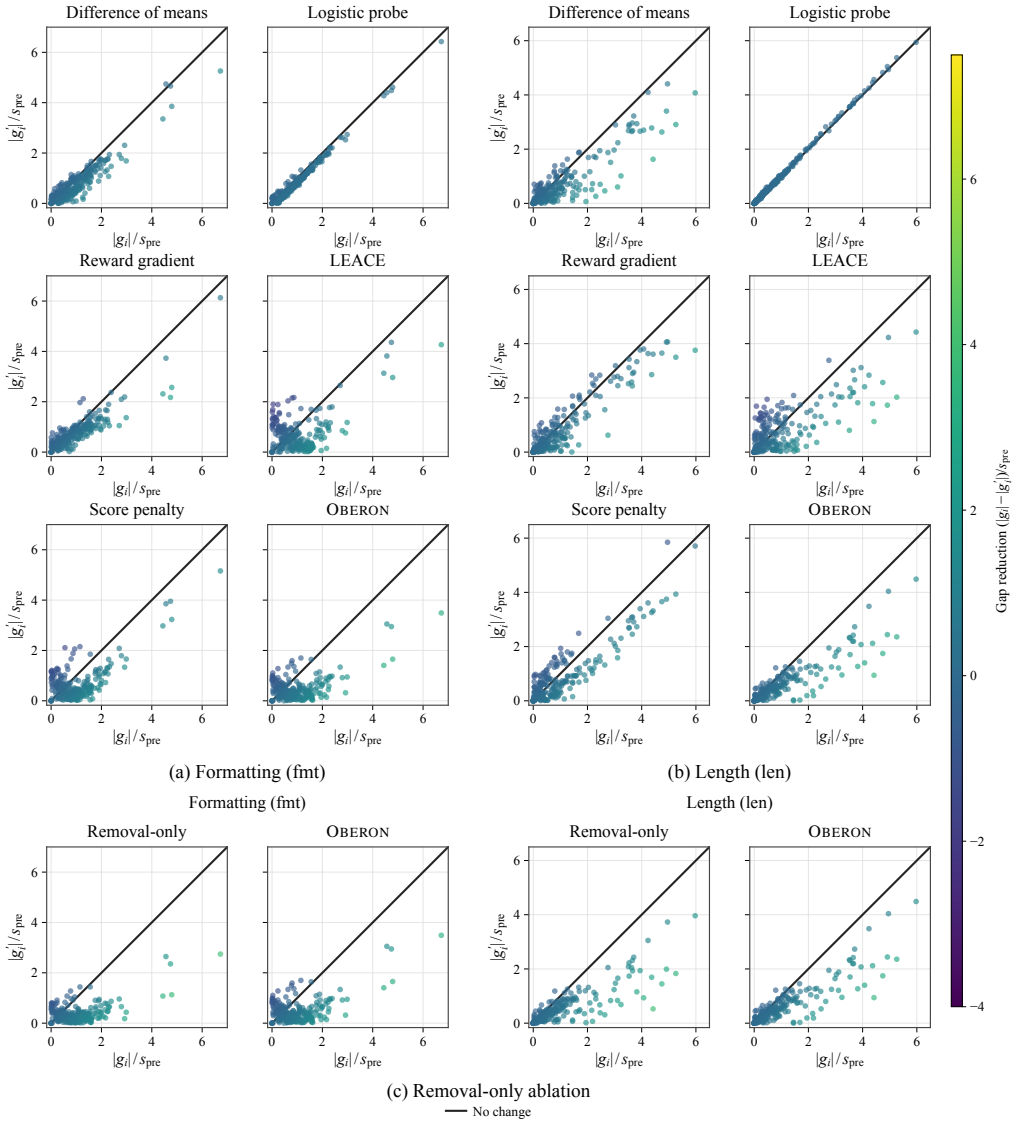


Figure 7: **Intervention effects on response pairs for Skywork-Reward-V2-Llama-8B.** Panels (a) and (b) compare OBERON with five baselines on `fmt` and `len`, respectively, while panel (c) compares OBERON with a variant trained using only the removal objective. The horizontal and vertical axes show absolute reward gaps before and after intervention, normalized by the mean absolute baseline gap s_{pre}. Color encodes $(|g_i| - |g_i'|)/s_{\text{pre}}$, with points below the diagonal indicating attenuation.

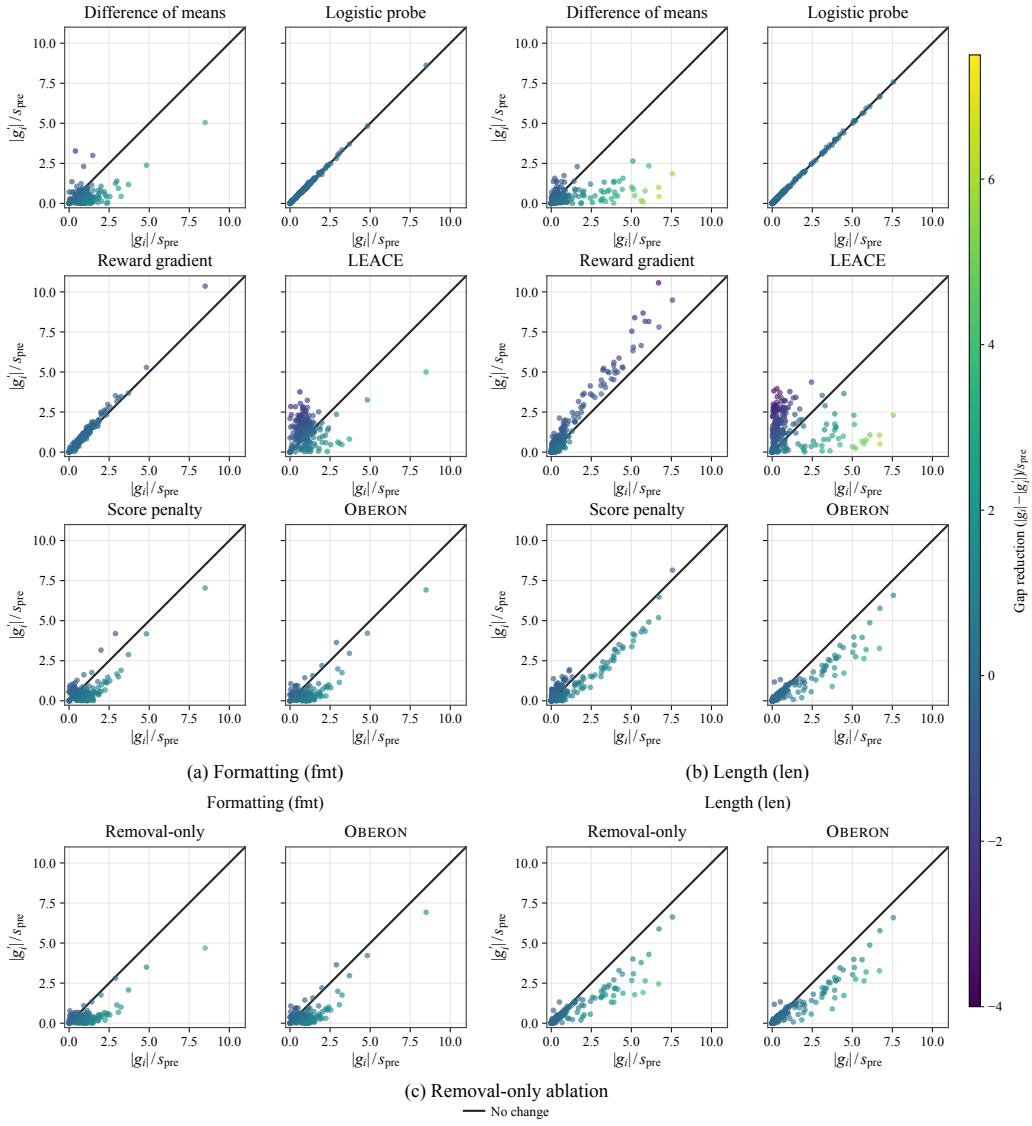


Figure 8: **Intervention effects on response pairs for Skywork-Reward-V2-Qwen3-8B.** Panels (a) and (b) compare OBERON with five baselines on `fmt` and `len`, respectively, while panel (c) compares OBERON with a variant trained using only the removal objective. The horizontal and vertical axes show absolute reward gaps before and after intervention, normalized by the mean absolute baseline gap s_{pre}. Color encodes $(|g_i| - |g_i'|)/s_{\text{pre}}$, with points below the diagonal indicating attenuation.

Table 7: **Reward gap attenuation relative to preference margin scale.** C_{pair} and C_{scale} report raw and normalized attenuation in percentages. s_{qual} is the ratio of the standard deviation of preference margins after intervention to that before intervention on the quality panel. Each column is aggregated separately. Controls are excluded because they use a different evaluation protocol.

	Formatting			Length		
Method	C_{pair} (%)	s_{qual}	C_{scale} (%)	C_{pair} (%)	s_{qual}	C_{scale} (%)
Difference of means	34.0	0.847	22.8	36.0	0.677	5.6
Logistic probe	10.4	0.995	10.1	−1.4	1.002	−1.3
Reward gradient	12.2	1.060	15.9	−3.3	1.017	−2.0
LEACE	37.7	0.902	26.0	9.6	0.766	−31.3
Removal-only	81.8	0.724	76.4	64.9	0.583	41.8
Score penalty	59.5	1.017	60.2	9.6	1.004	10.1
OBERON	72.8	0.880	69.2	40.9	0.798	25.5

Figure 10 relates original PDR to preference accuracy drop for each model on `fmt` and `len`. The model tables report OBERON's accuracy drops ranging from −0.30 to 1.33 percentage points on `fmt`, from 0.23 to 1.70 on `len`, from −0.30 to 0.80 on `syco`, and from 0.10 to 1.78 on `cstyle`. An average drop below one percentage point therefore does not imply the same bound for every model. On Skywork-Reward-V2-Qwen3-8B under `len`, OBERON achieves higher PDR than score penalty, at 60.9% versus 46.7%, but incurs a larger accuracy drop of 1.70 versus 0.27 percentage points. On the same model and condition, difference of means yields greater attenuation than OBERON, with C_{pair} of 56.2% versus 31.9%, but lower PDR of 58.4% and a larger accuracy drop of 11.87 percentage points. LEACE's average accuracy drop on `len` is also driven primarily by this model, with a drop of 30.13 percentage points compared with 0.63 to 0.97 on the other four models. These comparisons reveal trade-offs between joint effectiveness and collateral quality that average PDR alone does not capture, without establishing superiority at a matched quality cost.

C.2.4 REWARD-SCALE DIAGNOSTIC

We assess target-gap attenuation relative to the scale of chosen–rejected reward margins on a separate held-out quality panel. For each fit, let $q = \sum_i |g_i'| / \sum_i |g_i|$ denote the post-to-pre ratio of summed absolute target gaps, and let s_{qual} denote the post-to-pre ratio of margin standard deviations computed on the same quality pairs. We assume both denominators are positive and $s_{\text{qual}} > 0$. Under uniform reward rescaling $R' = aR$ with $a > 0$, we have $q = s_{\text{qual}} = a$. We therefore define scale-normalized attenuation as

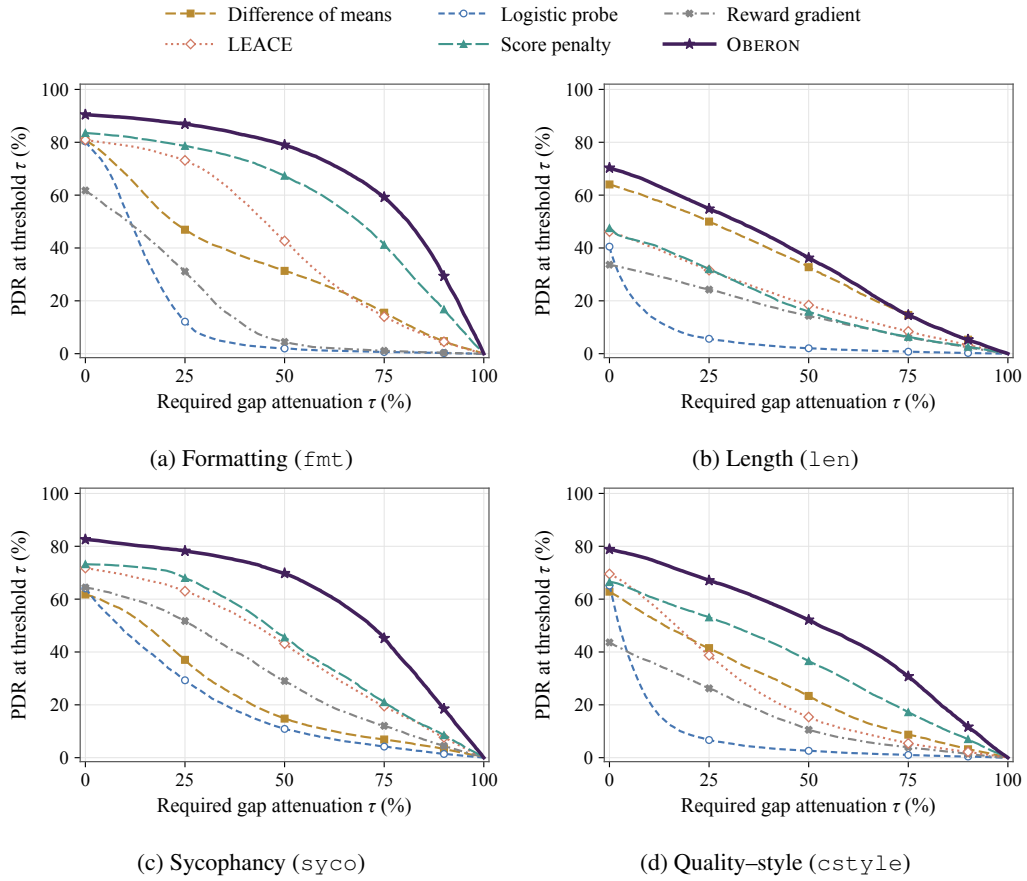
$$C_{\text{scale}} := 100 \left(1 - \frac{q}{s_{\text{qual}}}\right).$$

Table 7 reports raw and scale-normalized attenuation for seven methods on `fmt` and `len`. Each metric is computed per fit, summarized by the median over three draws within each model and split, and averaged equally across three splits and five models, covering 630 fits. The resulting mean C_{scale} for OBERON is 69.2% on `fmt` and 25.5% on `len`. On `len`, the mean raw and scale-normalized attenuation values are 36.0% and 5.6% for difference of means, and 9.6% and −31.3% for LEACE. At the fit level, $C_{\text{scale}} < 0$ means $q > s_{\text{qual}}$, which does not necessarily imply $q > 1$ or amplification of the raw target gaps. This diagnostic provides a reference-scale comparison rather than a decomposition of target-specific and global reward changes.

C.2.5 ABLATION STUDIES

We extend the ablation analysis to all five reward models on `fmt` and `len`, examining the objective weight and intervention layer. Both sweeps use $k = 10$ direction-training pairs, compared with $k = 40$ in the main experiments, and evaluate attenuation using C_{pair}. The weight sweep averages three split seeds, whereas the layer sweep uses one fit per layer. The results therefore characterize sensitivity under the ablation configuration rather than directly reproduce the main-table estimates.

Objective weight. Figure 11 reports the weight sweep from removal-only training at $\rho = 0$ to detection-only training at $\rho = 1$, with the intervention layer fixed at 12. Across all ten combinations



(a) Formatting (`fmt`)

(b) Length (`len`)

(c) Sycophancy (`syco`)

(d) Quality–style (`cstyle`)

Figure 9: **PDR under increasing reward-gap attenuation thresholds.** At threshold τ, PDR is the percentage of all test pairs with correct paired detection and a post-intervention reward gap satisfying $|g_i'| < (1-\tau)|g_i|$. The original PDR is recovered at $\tau = 0$. Curves average nine fit-level rates within each model, followed by an equal-weight average across five reward models. Detectors and interventions remain fixed across thresholds.

of model and condition, joint training at $\rho = 0.8$ achieves higher mean AUROC than removal-only training and greater mean attenuation than detection-only training. On `fmt`, mean attenuation at $\rho = 0.8$ ranges from 47.0% to 87.6%, compared with 0.5% to 19.2% at $\rho = 1$. On `len`, all five models retain positive mean attenuation at $\rho = 0.8$, whereas training at $\rho = 1$ yields negative mean attenuation despite higher mean AUROC than at $\rho = 0.8$. For Skywork-Reward-8B on `fmt`, increasing ρ from 0 to 0.8 raises mean AUROC from 0.823 to 0.963 while reducing mean attenuation from 93.4% to 87.6%. At $\rho = 1$, mean attenuation falls to 19.2%. Intermediate weights nevertheless exhibit trade-offs that vary across models. For Skywork-Reward-V2-Qwen3-8B on `fmt`, increasing ρ from 0.6 to 0.8 reduces mean AUROC from 0.912 to 0.900 and mean attenuation from 59.9% to 56.7%, while reducing the mean preference-accuracy drop from 0.17 to 0.00 percentage points. These results support combining the two objective terms, while showing that the balance among detection, attenuation, and collateral quality depends on the model and condition.

Intervention layer. Figure 12 reports the layer sweep with $\rho = 0.8$ and the direction fitted independently at each layer. The effects of layer choice vary across models, conditions, and metrics. On `fmt`, moving from layer 12 to 24 in Skywork-Reward-8B increases AUROC from 0.974 to 0.978 while reducing attenuation from 87.7% to 65.5% and the preference-accuracy drop from 1.0 to 0.0 percentage points. For Skywork-Reward-V2-Qwen3-8B, the same change increases attenuation from 52.9% to 59.5% but increases the preference-accuracy drop from 0.5 to 1.5 percentage points. On `len`, moving from layer 12 to 20 in Skywork-Reward-8B increases AUROC from 0.852 to 0.878 and attenuation from 33.8% to 56.3%, but increases the preference-accuracy drop from 1.0 to 4.0

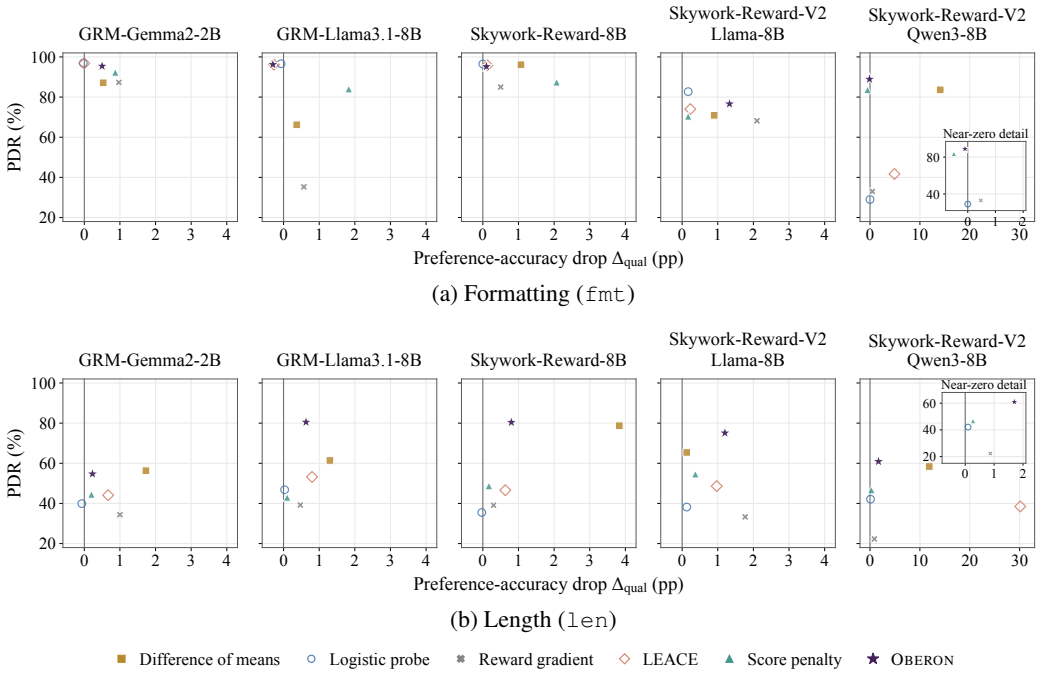


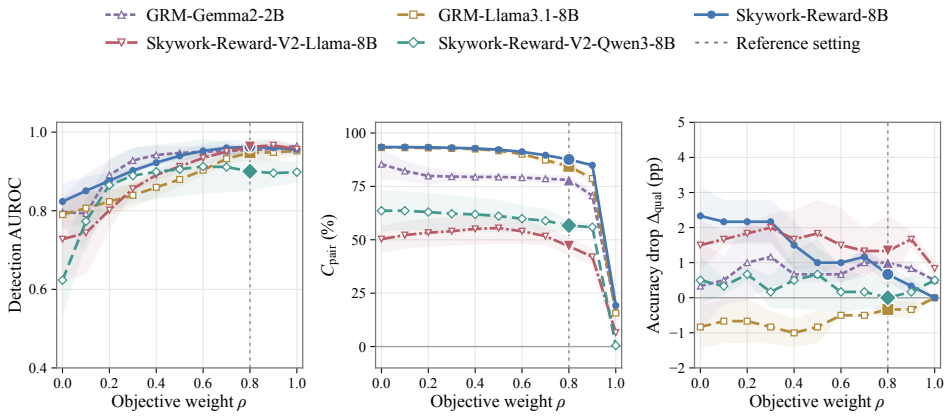
Figure 10: **Joint effectiveness and collateral quality across reward models.** Columns correspond to five reward models. Each point reports PDR and preference-accuracy drop for one method at its evaluated intervention setting. Accuracy drops are measured in percentage points, with positive values indicating degradation. All main panels share the PDR scale. The first four columns share a horizontal scale, while the Qwen3 column uses a wider range with insets resolving near-zero accuracy changes. These comparisons do not establish superiority at a matched quality cost.

percentage points. For Skywork-Reward-V2-Qwen3-8B on `len`, layer 16 yields better observed values on all three metrics than layer 12. AUROC increases from 0.826 to 0.848, attenuation increases from 26.2% to 29.8%, and the preference-accuracy drop decreases from 3.5 to 2.5 percentage points. These results indicate that layer 12 is a common reference setting rather than a uniformly optimal intervention location. Each combination of model, condition, and layer is evaluated with a single fit, so the stability of these comparisons across split seeds remains untested.

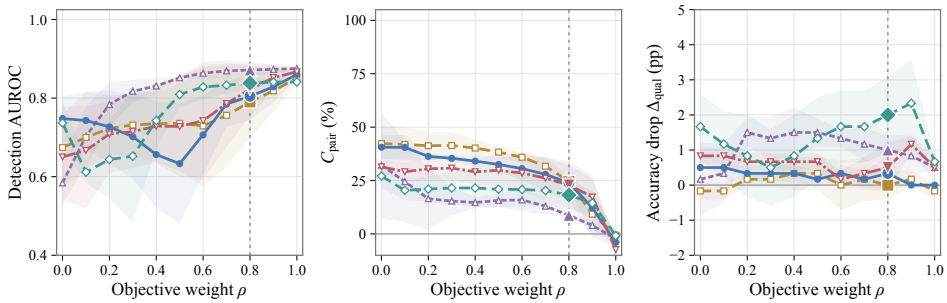
D LIMITATIONS AND FUTURE WORK

Coverage of evaluated contrasts. We evaluate OBERON on five reward models using four response contrasts drawn from existing benchmarks. While these contrasts provide a common basis for method comparison, they do not cleanly isolate undesirable reward biases in every case. In particular, the `fmt` and `len` variants have not been independently verified to preserve response quality, whereas paired responses differ in correctness for `syco` and in human preference labels for `cstyle`. As a result, a higher PDR does not always correspond to a desirable shift in reward preferences. To mitigate this limitation, we report collateral preference accuracy on separate evaluation panels and apply the scale diagnostic to `fmt` and `len`. Nevertheless, broader evaluation with independently annotated quality judgments, as well as comparisons under matched quality trade-offs, would enable a more precise assessment of the benefits of intervention.

Generalization to policy optimization. Our experiments focus on changes in reward scores and pairwise preference ordering over fixed response pairs. This setup enables controlled comparison of intervention effects without the confound of a changing generation policy. However, it does not capture downstream settings in which the edited reward model is used to rank newly generated candidates or to optimize a policy directly. An important direction for future work is to evaluate best-



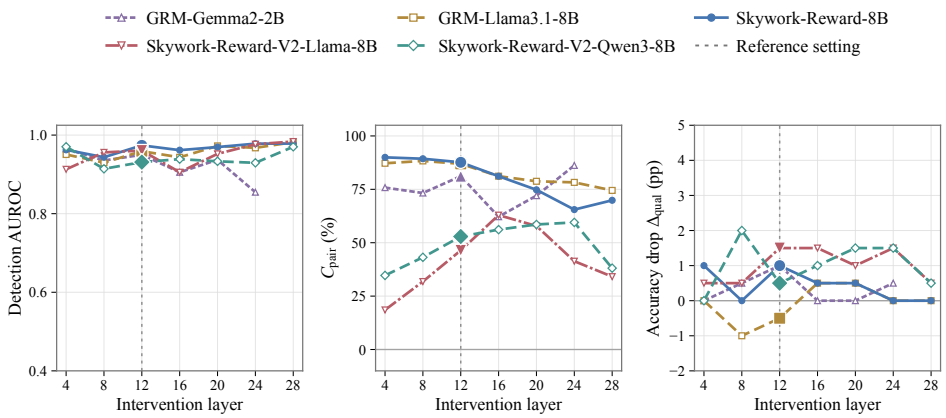
(a) Formatting (`fmt`)



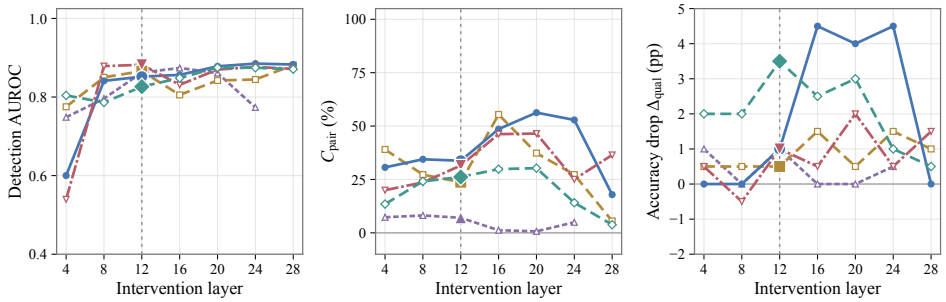
(b) Length (`len`)

Figure 11: **Sensitivity to the objective weight.** Results for OBERON on formatting and length across five reward models at layer 12. Curves show means across three split seeds, with shaded bands denoting $\pm$ one standard deviation. Dashed vertical lines and enlarged markers indicate $\rho = 0.8$.

of-N selection and reward-guided policy training, in order to determine whether the improvements persist once the edited reward model begins to influence the response distribution itself.



(a) Formatting (`fmt`)



(b) Length (`len`)

Figure 12: **Sensitivity to the intervention layer.** Results for OBERON on formatting and length across five reward models, with $\rho = 0.8$. Each point represents one fit with split seed 1 and draw 0, with the direction refitted at each layer. Dashed vertical lines and enlarged markers indicate layer 12. Layer indices are zero-based, and layer 28 is unavailable for the 26-layer GRM-Gemma2-2B.