DACO: DISCRIMINABILITY-AWARE CONTENT OPTIMIZATION FOR AI-GENERATED CONTENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model (LLM) evaluators increasingly rank and score AI-generated or AI-rewritten text. When powerful generative models rewrite content at scale, different texts can become more similar in wording, organization, and presentation. We call this effect content homogenization. We ask whether homogenization changes the internal features used by an evaluator even when its accuracy remains similar, and whether features that remain important after homogenization can guide content optimization. We study these questions through representation analysis using Principal Component Analysis (PCA) components of pairwise hidden-state differences to identify features that predict evaluator preferences. We evaluate three open-weight evaluators (Gemma 4 12B, Ministral 3 14B, and Phi-4 14B) across human-preference judgments, assistant-response ranking, and hiring. We further test generalization with Llama 4 Scout (109B total, 17B active), Llama-3.1-405B-Instruct, Gemini 3.6 Flash, with an additional blind human evaluation. Homogenization changes evaluator accuracy by only about one percentage point while substantially reorganizing which PCA components predict evaluator decisions. This shows that stable accuracy can hide changes in the features associated with evaluator decisions. Held-out functional validation further shows that components that remain important after homogenization can provide useful guidance for content optimization. Based on this finding, we introduce Discriminability-Aware Content Optimization (DACO), a representation-guided content optimization method that uses validated guidance for a single guided rewrite while preserving source content. DACO has zero protected-literal failures while preserving source content. These results show that representation analysis can reveal changes in evaluator features that accuracy alone misses, while also providing a practical basis for content optimization.

1 INTRODUCTION

LLM evaluators (large language models used to compare, rank, or score text) are increasingly used for tasks that rely on human judgment Zheng et al. (2023); Liu et al. (2023). In these evaluation settings, a powerful generative model often rewrites the text before evaluation. For example, powerful generative models (e.g., ChatGPT, Claude, Gemini etc) may polish resumes, assistant answers, and applicant profiles before evaluation. When many texts pass through similar generative models, they tend to look more alike. Recent studies similarly show that LLMs reduce the diversity of the content they produce Doshi & Hauser (2024); Moon et al. (2025); Sourati et al. (2026). We call this effect content homogenization: the tendency of powerful generative rewriting to make different texts more similar. For evaluation, the key issue is that evaluators depend on differences between items. When rewriting smooths out the features that once separated a stronger item from a weaker one, the features used to rank them can change. We therefore study how content homogenization changes the features that predict an LLM evaluator's preference or ranking decision. We study this as LLM evaluation under generative distribution shift (i.e., evaluation of texts reshaped by generative rewriting). We use hiring as a high-stakes application to test whether the same representation shift appears outside standard human-preference data.

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.

Behavioral measurements show that a decision changed, but not which internal features predict it. To examine this internal change, we analyze the evaluator's hidden representations and identify the PCA components that predict its decisions before and after homogenization. Changes in the predictive importance of these components show whether the features used by the evaluator change even when behavioral differences are modest. We call this phenomenon homogenization-induced representation shift and test its causal and cross-model consistency. For content optimization, we build on this analysis with DACO (Discriminability-Aware Content Optimization), a representation-guided content optimization method. DACO derives guidance from validated components that remain important after homogenization and uses it for single guided rewrite while preserving meaning and factual content.

These two parts of the study, understanding homogenization-induced representation shifts and using them for content optimization, lead to four research questions:

- **RQ1 (Effect of homogenization).** How does content homogenization affect the discriminative performance (the evaluator's ability to separate a stronger item from a weaker one) and the predictive structure of internal representations in LLM evaluators?
- **RQ2 (Causality and cross-model consistency).** Are homogenization-induced representation shifts causally involved in evaluator decisions, and are they consistent across LLM families and model scales?
- **RQ3 (DACO effectiveness).** Can targeting PCA components that remain important after homogenization match or outperform strong content-optimization alternatives while preserving content quality?
- **RQ4 (Generalization and external validity).** Do the observed representation shift and DACO generalize across human-preference, assistant-response, and hiring domains, to a closed black-box evaluator and to human judgment?

We study three open-weight evaluators, Gemma 4 12B Team et al. (2026), Ministral 3 14B Liu et al. (2026), and Phi-4 14B Abdin et al. (2024), across three domains: human-preference judgments (SHP, Ethayarajh et al. (2022)), assistant-response ranking (OASST1, Köpf et al. (2023)), and hiring (Djinni, Drushchak & Romanyshyn (2024)). Across these settings, homogenization leaves evaluator accuracy largely unchanged while reorganizing the PCA components that predict evaluator decisions. This shows that similar evaluation accuracy can hide a change in the features used by the evaluator. We further find that components that remain important after homogenization can provide useful guidance for content optimization. DACO uses this validated guidance in a single guided rewrite while preserving source content and remaining competitive with the comparison methods.

Contributions.

- We characterize how content homogenization changes LLM evaluator discriminability and the predictive importance of PCA components, across three evaluator families (Gemma, Ministral, Phi-4) and three datasets representing distinct evaluation domains: SHP (human-preference judgments), OASST1 (assistant-response ranking), and Djinni (hiring).
- We causally test selected PCA components with decreased, preserved, or increased predictive importance against matched random orthogonal controls and report the observed intervention effects.
- We propose DACO, a representation-guided content optimization method that targets validated PCA components that remain important after homogenization, and we evaluate it against five comparison conditions (B0, B1, B2, A1, A2) under matched factual-preservation constraints.
- We test generalization across three domains (SHP, OASST1, Djinni), a second rewriting model (Llama 4 Scout), a supplementary white-box evaluator (Llama-3.1-405B-Instruct, Grattafiori et al. (2024)), a closed black-box evaluator (Gemini 3.6 Flash, Google (2026)), and blind human evaluation.

2 Related Work

LLM-as-a-judge (using an LLM to score or compare outputs) is now common for ranking and preference evaluation Zheng et al. (2023); Liu et al. (2023), with dedicated fine-tuned judge models also

proposed as scalable evaluators Zhu et al. (2025). Our study examines how open-weight LLM evaluator decisions shift when the judged content is first rewritten by a powerful generative model. LLM evaluators already respond to factors beyond content quality. They show position bias (a tendency to favor a text based on whether it appears first or second) Shi et al. (2025), and their judgments can be inconsistent across settings Wang et al. (2024). We control for position bias and examine content homogenization as another systematic source of change in evaluator behavior.

Prior work reports reduced diversity in model-assisted and model-generated content, including creativity Doshi & Hauser (2024), human versus model writing Moon et al. (2025), and linguistic diversity at scale Sourati et al. (2026). These studies characterize the produced text; we instead examine how homogenization affects the evaluator that judges it. Related representation-analysis methods include representation engineering Zou et al. (2023), activation patching for causal testing of internal components Zhang & Nanda (2024), centered kernel alignment (CKA) for comparing representation geometry Kornblith et al. (2019), and RankSteer for steering evaluator activations to change rankings Wang et al. (2026). We use these methods to identify and test homogenization-induced representation shifts; unlike activation steering, DACO uses validated PCA components to guide content rewriting rather than modifying evaluator activations. Related work also optimizes text to change LLM rankings. Strategic text sequences add tokens to increase a target product's visibility Kumar & Lakkaraju (2024). RAF (Rank Anything First) searches for ranking-improving text through two-stage token optimization and is the strongest published competitor for the hiring task Xing et al. (2026). In AI-assisted hiring, prior work studies strategic manipulation and fairness Cohen et al. (2025) and prompt injection (inserting instructions into the input to change the model's output) in automated resume screening Baxi et al. (2026). Unlike direct ranking optimization, DACO starts from the effect of content homogenization on evaluator representations and uses PCA components whose predictive importance is preserved or increased after homogenization to guide content rewriting. We compare DACO with RAF on the Djinni hiring task.

3 METHODOLOGY

This section presents the methodology in the execution order shown in Figure 1. (1) **Natural Inputs** are paired texts drawn from three datasets: SHP (each pair contains two human-written responses with a human preference label), OASST1 (each pair contains two AI assistant responses to the same prompt with a ranking), Djinni (each pair compares a target applicant profile with a competitor profile for the same job description). These original texts form the natural condition and are given to (2) the **Homogenizer** (gpt-oss-120b in the primary experiments; Llama 4 Scout in the second-rewriting-model evaluation). The homogenizer rewrites each text to make the paired texts more alike while preserving factual content. The resulting text x_H forms the homogenized condition (and later serves as B0) (3) then **LLM Evaluators (E)** (Gemma 4 12B, Ministral 3 14B, and Phi-4 14B) compare the two texts in each pair and evaluate them in both presentation orders (A/B and B/A; A and B are the two texts in the pair). Evaluating both orders controls position bias. For each evaluation, we save the preference score and the hidden representations at a fixed token position. Using these saved representations, (4) In **Pair-difference (d_i)** stage computes the difference between the two texts' hidden representations within each presentation order and calculates the averages of those two differences. This produces one representation for each pair. (5) **PCA + Ridge Surrogate** stage identifies the PCA components that predict the evaluator's preference margin (how strongly the evaluator favors one text over the other). (6) Next, **Importance Shift (ΔPI_k)** compares each component's predictive importance between the natural and homogenized conditions to identify whose importance decreases, increases, or remains preserved after homogenization. The strongest components from this analysis then enter (7) **Causal Validation**, which tests whether changing those components affects the evaluator's decision (using rescue or ablation interventions against matched random orthogonal controls). In addition, Llama-3.1-405B-Instruct is tested on a smaller subset, with causal intervention only when its held-out surrogate passes the same validity gate, to evaluate generalization at large scale. Selected components whose importance remains preserved or increases enter (8) **Functional Validation**, which converts each component into written guidance and validates it on held-out examples (to determine whether the guidance moves the representation in the intended direction). (9) Guidance that passes this test becomes **Validated Guidance**. The guidance for the component that passes functional validation continues along the main path to DACO, while (9a) the broader validated guidance set is grouped separately into the **Practical Insights** for interpreting recurring actionable patterns. The main path then continues to (10) **DACO Rewrite**, which applies

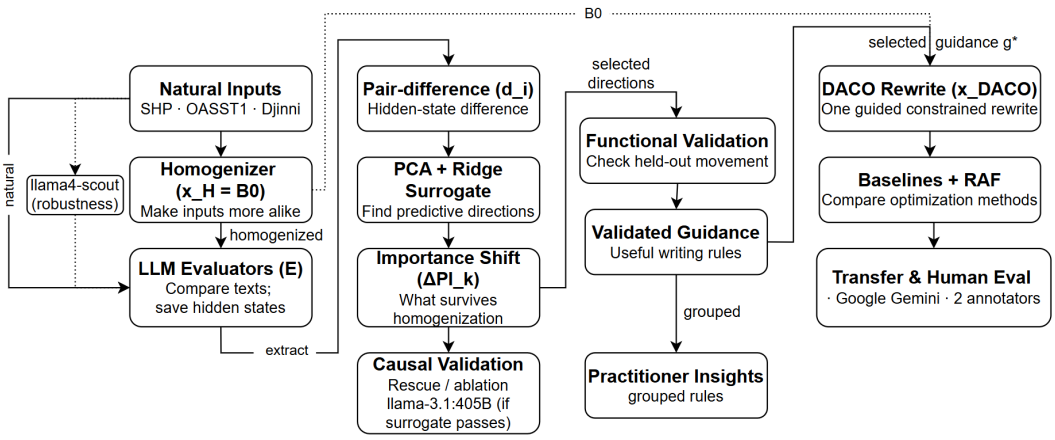

Figure 1: Complete experimental pipeline. The pipeline homogenizes natural texts, evaluates them with three open-weight LLM evaluators, and analyzes their representations to derive validated guidance for DACO. Additional evaluations test robustness, transfer, and human judgment.

the validated guidance once to the homogenized text while preserving factual content. The resulting DACO rewrite enters (11) **Baselines + RAF**, where it is compared with B0, B1, B2, A1, A2, and RAF. RAF (Rank Anything First) is a separate adversarial ranking baseline that uses two-stage token optimization to alter an LLM ranking. Finally, (12) **Transfer & Human Evaluation** evaluates the transfer to the closed black-box evaluator and through blind human evaluation. The behavioral and representation analyses [Stages 1–6] address RQ1; causal and consistency analyses [Stage 7] address RQ2; functional validation and the DACO comparisons [Stage 8–11] address RQ3; and the additional domains [Stages 1–11], additional models [Stages 2 and 7], and the closed black-box evaluator and human evaluation [Stage 12] address RQ4.

3.1 Representation readout and analysis

The evaluation records two things for each pair at a fixed token position: the preference score and the hidden representation. We place the same `END CANDIDATE` marker after each text and read the hidden representation at the last token of that marker, so the token position is the same for every text. We run each pair in both orders, A, B and B, A, and map the outputs back to the original text identities; running both orders controls position bias Shi et al. (2025). We then average the two orders. The pair-difference representation is

$$d_i^{AB} = h_A^{AB} - h_B^{AB}, \qquad d_i^{BA} = h_A^{BA} - h_B^{BA}, \qquad d_i = \frac{d_i^{AB} + d_i^{BA}}{2}. \tag{1}$$

The pair-difference vector d_i is the difference between the two texts' hidden states, averaged over both orders.

For the preference score, we use the evaluator's next-token probability for single-token labels, and the full label-sequence log-likelihood difference otherwise. The order-corrected preference margin is

$$m_i = \log P(A \mid x_i) - \log P(B \mid x_i), \tag{2}$$

after mapping both orders to the original text identities and averaging. This log-probability difference is the log ratio $\log(P(A \mid x_i)/P(B \mid x_i))$, where the sign indicates the preferred text and the magnitude indicates the strength of the relative A/B preference. We orient the margin by the preference label,

$$s_i = y_i\, m_i, \qquad y_i = +1 \text{ if } A \text{ is preferred, } -1 \text{ otherwise,} \tag{3}$$

so $s_i > 0$ means the evaluator agrees with the preference label, and a larger s_i means stronger agreement with the preference label. To measure the effect of homogenization on behavior, we compare s_i between the natural and homogenized conditions with preference accuracy, the signed margin, paired bias-corrected and accelerated (BCa) bootstrap 95% confidence intervals Efron (1987), and the exact McNemar test for the paired accuracy change.

We then measure the change in the size (norm) of the pair-difference vector. For each pair at layer ℓ we take the ratio of the homogenized norm to the natural norm,

$$C_i^{(\ell)} = \frac{\|d_{i,H}^{(\ell)}\|_2}{\|d_{i,N}^{(\ell)}\|_2}, \tag{4}$$

and report the median ratio with a paired BCa 95% confidence interval. We relate the norm change to behavior with a Spearman correlation between C_i and the signed-margin change $\Delta s_i = s_{i,H} - s_{i,N}$.

To find which PCA components predict the evaluator's preference margin, we fit one shared PCA basis from both conditions and project both conditions into it. In each condition we fit a ridge surrogate that predicts the margin from the PCA coordinates,

$$m_N \approx \beta_N^\top z_N, \qquad m_H \approx \beta_H^\top z_H. \tag{5}$$

We select the PCA dimension K on validation data by margin-prediction performance (maximizing the lower of the natural and homogenized validation R^2 values), with the test set excluded, and the final analysis uses $K = 512$ for all three evaluators. We report the held-out surrogate R^2 and interpret component importance only when both conditions have $R^2 > 0$ and the lower bound of the 95% bootstrap interval for improvement over the constant-mean baseline is above zero. We score each component by held-out permutation importance (the drop in the surrogate's prediction performance when one PCA coordinate is shuffled) Fisher et al. (2019) and compare conditions by

$$\Delta PI_k = PI_{k,H} - PI_{k,N}, \tag{6}$$

with paired BCa intervals. We classify each of the 512 components as decreased, increased, preserved, or unsupported. A component is decreased if $\Delta PI_k < 0$ and increased if $\Delta PI_k > 0$, in both cases with a paired BCa interval that excludes zero. A component is preserved when neither change is supported and its homogenized-condition permutation-importance 95% BCa interval is above zero, and unsupported otherwise. We summarize the overall change in the importance ranking with Kendall τ_b between the natural and homogenized rankings, with a 95% bootstrap confidence interval, and report a ranking reversal only when the predictive-importance ordering clearly reverses.

3.2 Causal and cross-model validation

We test whether the selected components affect the evaluator's decision using causal interventions inside the model. We select the strongest supported components (largest supported ΔPI_k decreases or highest homogenized permutation importance) on training and validation data, at most two components with decreased importance and two with preserved or increased importance per evaluator, and apply the interventions at the same layer and token position used for the pair-difference representation Zhang & Nanda (2024). We use two interventions: rescue and ablation. Rescue restores the reduced contribution of a component with decreased importance. Ablation removes the remaining contribution of a component with preserved or increased importance. The exact intervention equations are in Appendix A. We continue the forward pass, measure the change in the evaluator's A/B margin, and run both orders with A/B label remapping. For each selected component we run the same intervention on a matched random orthogonal vector, which separates a component-specific effect from a generic perturbation. We aggregate the effects with paired BCa intervals and apply Benjamini-Hochberg false discovery rate (BH-FDR) correction across the tests at $q < 0.05$ Benjamini & Hochberg (1995).

We also test whether the homogenization effect is consistent across evaluators. Because PCA components are specific to each model, we compare the homogenization-induced shift rather than numbered components. For each evaluator M we stack its natural and homogenized pair representations into $H_{M,N}$ and $H_{M,H}$, form the shift $D_M = H_{M,H} - H_{M,N}$, and compare shift geometry across evaluator pairs with linear centered kernel alignment (CKA) Kornblith et al. (2019). We compare per-example signed-margin changes across evaluators with Spearman correlation.

3.3 Functional validation and guidance

We derive written guidance from a selected component, then test whether it changes the component projection on unseen examples. We call this functional validation. We use Ministral 3 14B for guidance generation. For DACO, we select up to two preserved or increased-importance components

with the highest homogenized permutation importance. For a component whose importance is preserved or increased, we rank the training examples by their projection on the component, $r_i = z_i^\top v$, take the highest and lowest groups, and give both groups to the rewriting model (gpt-oss-120b). We ask it for one short, actionable guidance statement that separates the two groups in plain language, without referring to components, vectors, or scores. We then validate the statement on held-out validation examples that were not used to write it. For each example we apply the guidance in a guided rewrite and measure the change in its component projection,

$$\Delta_i = (z_i^{\text{after}} - z_i^{\text{before}})^\top v, \qquad \text{mean projection movement} = \frac{1}{n}\sum_{i=1}^{n} \Delta_i. \tag{7}$$

We keep a statement only when its mean projection movement is strictly positive and it stays short, actionable, and free of any reference to PCA, vectors, activations, models, or scores. Each statement is validated on $n = 12$ held-out examples per component, and the generation settings are in Appendix A. This step produces the single validated guidance statement that DACO uses and the broader validated set used for the practical insights. For the Practical Insights, we pass only the validated statements to gpt-oss-120b and group them conservatively when they give materially the same advice, with all component identifiers, scores, ranks, and labels removed.

3.4 DACO AND COMPARISON CONDITIONS

DACO derives guidance from a selected component that remains important after homogenization, validates the guidance on held-out representations, and applies a single guided rewrite to the homogenized text. The next evaluation measures both the preference and the projection change on the selected component. The rewrite uses the same factual-preservation constraints and adds no new claims, numbers, or entities. Appendix A gives the complete step-by-step procedure.

We compare DACO against five conditions built from the same original pairs. B0 is the homogenized text from the homogenization step. B1 is Best-of-4: it reuses B0 and adds three generic fact-preserving rewrites, then uses the Ministral evaluator to pick the best of the four; its sampling settings are in Appendix A. B2 is a homogenization-aware prompt that tells the rewriting model the content will be judged in an AI-homogenized setting and asks for more distinctiveness while preserving content, with no representation information. A1 uses PCA-component guidance learned from the natural condition, without selecting components by their importance change under homogenization. A2 uses the same steps with matched random vectors in place of the selected components. These five conditions isolate the roles of generic rewriting (B0), extra sampling and search (B1), homogenization awareness (B2), generic representation guidance (A1), and component selection (A2). Generation settings and preservation rules are the same across the single-rewrite conditions.

For the hiring task we use a fixed pool of the target applicant and four competitors. Only the target applicant's text changes across B0, B1, B2, A1, A2, and DACO; the job and the four competitors stay fixed. We compare the target against each competitor in both orders and report the pairwise win rate as the fraction of the four competitors beaten. We also compare DACO with RAF (Rank Anything First), which changes LLM ranking through two-stage token optimization and is the strongest published adversarial ranking baseline for this task Xing et al. (2026). For RAF, we report its ranking result and optimized prompt length; RAF appends the optimized prompt to the source text rather than rewriting the source content. RAF run settings are in Section 4 and Appendix A.

4 EXPERIMENTAL SETUP

Datasets and splits. Stanford Human Preferences (SHP) provides 600 human-written preference pairs across domains and is the primary representation-analysis dataset Ethayarajh et al. (2022). OpenAssistant OASST1 provides 300 ranked same-prompt response pairs for the assistant-response domain Köpf et al. (2023). The Djinni recruitment dataset provides 100 Product Manager and 100 UI/UX applicant profiles with matched job descriptions for the hiring application Drushchak & Romanyshyn (2024). We split each dataset into training, validation, and test: SHP 360/120/120, OASST1 180/60/60, and Djinni 100/40/60, keeping the Product Manager and UI/UX balance in each Djinni split. All model, layer, PCA dimension, component, and guidance choices use only the training and validation data, and the test set is held out from these choices. A deterministic screen keeps

only meaningful natural-language fields (excluding URL-only, machine-like, or symbol-dominant fields); ineligible fields advance to the next eligible item in the same dataset, split, and stratum, preserving the sampling structure.

Homogenization and factual preservation. For each source text we produce a homogenized version with gpt-oss-120b at temperature 0 under the factual-preservation constraints Agarwal et al. (2025). This homogenized version is also the B0 condition; we generate it once and reuse it. We enforce protected-literal preservation mechanically by replacing source digits, URLs, emails, and other literal-bearing spans with opaque placeholders before generation. The output must keep these placeholders in source order and introduce no model-generated digits, and we then restore the original literals and validate them exactly. We call any violation of these protected literals a protected-literal failure. The same preservation procedure applies to every generated condition, so differences between conditions come from their guidance or search, not from different preservation rules. We use term frequency-inverse document frequency (TF-IDF) cosine similarity as a diagnostic text-similarity check; it does not determine rewrite validity.

Evaluators and infrastructure. The three primary evaluators are Gemma 4 12B Team et al. (2026), Ministral 3 14B Liu et al. (2026), and Phi-4 14B Abdin et al. (2024), each run with its official chat template on the SPHERE research infrastructure Mirkovic et al. (2024). A duplicate Gemma run checks that the evaluator scores and pair representations reproduce exactly.

Second rewriting model. To test whether the effect depends on the writing style of gpt-oss-120b, we regenerate the homogenized condition with Llama 4 Scout Meta (2025). On 60 SHP, 30 OASST1, and 30 Djinni items (120 total), we recompute the RQ1 and RQ2 measures across the three primary evaluators.

Large-scale white-box evaluator. We extend the representation analysis to Llama-3.1-405B-Instruct Grattafiori et al. (2024) on a subset of 60 SHP, 30 OASST1, and 30 Djinni items (120 total). We test natural versus homogenized discriminability, representation-norm change, and predictive-importance shift. Component interpretation and causal intervention use the same held-out surrogate validity criterion as the primary evaluators; access and implementation details are in Appendix A.

Closed black-box evaluator. We evaluate with the closed black-box evaluator Gemini 3.6 Flash, accessed only through its API Google (2026). Selection rate is the fraction of pairs for which the target applicant in Djinni, or Candidate A in SHP and OASST1, is selected in both orders after A/B label remapping; AB/BA order consistency records whether the same text identity is selected across orders. Both rates use $n = 420$ paired evaluations per method; API settings are in Appendix A.

Human evaluation. Two independent annotators blindly evaluate 180 unique samples, 60 from each of SHP, OASST1, and Djinni. We label B0 and DACO as Candidate A and Candidate B and counterbalance their positions across annotators (Appendix A). Each annotator gives a pairwise judgment (A better, Tie, or B better) and an independent preservation judgment (Preserved or Not preserved) for each candidate. We report linearly weighted Cohen's kappa and raw agreement for preference, and Cohen's kappa and raw agreement for preservation, computed before any aggregation. For aggregation, matching judgments are retained, a directional judgment paired with Tie stays directional, and opposing A/B judgments remain unresolved; no third adjudicator is used.

RAF protocol. RAF follows its published protocol on 60 Djinni test targets with Phi-4; full settings are in Appendix A.

5 RESULTS

RQ1: Effect of Homogenization. Table 1 shows that preference accuracy stays close to its natural level, moving by about one percentage point across evaluators, while all signed-margin 95% BCa intervals include zero. The duplicate-Gemma control passed its score and representation consistency check, and the held-out representation surrogate passed its validity criterion for all three evaluators. Representation-norm ratios are above 1 for all three evaluators, so we report a representation-norm change rather than compression. The association between representation-norm change and signed-margin change is weak across evaluators (Spearman $|\rho| \leq 0.191$; $n = 180$ each). Positive Kendall τ_b values show partial rank stability with substantial predictive-importance reordering, consistent with a homogenization-induced representation shift.

Table 1: Homogenization effects across the three primary evaluators.

Evaluator	Pref. accuracy Natural → Homog.	Δ signed margin [95% BCa CI]	Rep.-norm ratio [95% CI]	Kendall τ_b [95% CI]
Gemma	63.33% → 62.22%	−0.0854 [−0.5480, 0.3266]	1.0456 [1.0315, 1.0656]	0.2336 [0.1588, 0.2675]
Ministral	66.11% → 66.67%	−0.0672 [−0.3481, 0.2426]	1.0230 [1.0092, 1.0373]	0.3397 [0.2188, 0.3501]
Phi-4	67.22% → 66.11%	−0.3045 [−0.7837, 0.1599]	1.0134 [1.0012, 1.0313]	0.3283 [0.2192, 0.3308]

Table 2: Paired signed-margin differences between DACO and comparison conditions with 95% BCa CIs.

Evaluator	B0	B1	B2	A1	A2
Gemma	−0.3257 [−0.6617, −0.0517]	−0.2023 [−0.5862, 0.1797]	−0.0802 [−0.3297, 0.1704]	−0.0258 [−0.4478, 0.4482]	+0.1382 [−0.2745, 0.4996]
Ministral	−0.0979 [−0.3007, 0.0805]	−0.0825 [−0.3236, 0.1521]	+0.0095 [−0.1509, 0.1788]	+0.0703 [−0.1446, 0.3002]	+0.1290 [−0.1188, 0.3582]
Phi-4	+0.0542 [−0.1812, 0.2714]	+0.1434 [−0.1944, 0.5053]	−0.1069 [−0.3173, 0.1058]	+0.1038 [−0.2170, 0.4339]	+0.2500 [−0.0382, 0.5359]

Every evaluator carries PCA components with decreased, increased, and preserved predictive importance (Appendix Table 9), so homogenization does not simply raise or lower importance across the board. Appendix Figure 2 shows moderately positive behavioral agreement across evaluator pairs (Spearman $\rho = 0.3936$–0.4402), while representation-shift similarity varies more widely (linear CKA $= 0.1455$–0.4083), strongest for Phi-4–Ministral.

RQ2: Causality and Cross-Model Consistency. Across 12 rescue and ablation tests against matched random orthogonal controls, none reached BH-FDR significance at $q < 0.05$. In the supplementary Llama-3.1-405B-Instruct evaluation, the held-out surrogate failed the validity criterion ($R^2 < 0$ in both natural and homogenized conditions). We therefore base the component evidence on the statistical selection and the held-out functional validation.

One of the two selected DACO components passed functional validation: PC2 (Insight 1 in Table 3), with a predictive-importance increase of 0.0692 (95% BCa CI [0.0149, 0.1256]) and held-out mean projection movement 0.6047 over $n = 12$ examples. Applying the same procedure to 96 preserved or increased-importance Ministral components produced 34 functionally validated guidance statements, conservatively grouped (using gpt-oss-120B) into the five practical insights in Table 3. Insight 1 corresponds to PC2; the other four are separate semantic groups without PC2 support.

RQ3: DACO Effectiveness. DACO preference accuracy is 62.78% on Gemma, 63.89% on Ministral, and 66.11% on Phi-4. It matches or exceeds the other conditions in preference accuracy in 5/5, 3/5, and 4/5 comparisons, respectively. The paired analysis yields 0 strictly positive 95% BCa intervals and 0 exact McNemar tests with $p < 0.05$; Gemma versus B0 is the only signed-margin interval entirely below zero (-0.326 [-0.662, -0.052]). Absolute accuracies and Djinni win rates are in Appendix Tables 11 and 12. DACO uses a single guided rewrite, whereas B1 uses Best-of-4 sampling and selection.

DACO generation had zero protected-literal failures across 1100 rows. On the paired Phi-4/Djinni evaluation, DACO achieved a 47.50% pairwise win rate versus 56.67% for RAF. RAF used a 10-token optimized prompt, and DACO used a single guided rewrite under the study's factual-preservation rules; full RAF ranking statistics are in Appendix A.

RQ4: Generalization and External Validity. With Llama 4 Scout, 113 of 120 targeted items completed all required rewrites (Appendix B). On Gemini 3.6 Flash, selection rates range from 36.90% to 41.43% and order consistency from 77.86% to 84.76% (Table 4, 2520 pair-level metrics). DACO

Table 3: Practical insights from 34 validated statements.

#	Practical insight
1	Include concrete, quantitative, results-oriented details (metrics, specific achievements, measurable outcomes).
2	Provide thorough, detailed answers instead of short, generic ones.
3	Keep your writing concise; use brief, clear sentences and omit unnecessary detail.
4	Give a concise, focused response that directly answers the prompt, either by asking a clarifying question or providing a brief answer.
5	Give answers that are brief but include all essential details.

Table 4: Gemini 3.6 Flash evaluation; $n = 420$ paired evaluations per method.

Method	Selection	AB/BA consistency
A1	36.90%	77.86%
A2	41.43%	83.33%
B0	40.00%	84.05%
B1	40.00%	82.14%
B2	41.19%	84.76%
DACO	39.76%	82.62%

has 39.76% selection and 82.62% consistency, both within these observed ranges. Two annotators judged 180 blind samples. Pairwise preference agreement was 30.56% (weighted Cohen's kappa 0.0197). After deterministic aggregation, final selections were 68 B0, 58 DACO, 53 Tie, and one unresolved. Preservation agreement was 99.44% for both candidate positions, substantially higher than preference agreement.

6 Discussion and Limitations

The central finding is that content homogenization changes LLM evaluator accuracy little while reorganizing the PCA components that predict evaluator decisions. This means that stable accuracy can hide a change in the features used by the evaluator. The weak relation between representation-norm change and signed-margin change further shows that behavioral and representation analyses capture different parts of the effect. Held-out functional validation supports the practical use of the identified components for guidance, while the causal interventions did not establish direct causal effects under the current protocol. The Llama-3.1-405B-Instruct analysis also stopped at the predefined surrogate-validity gate, so we do not extend the causal interpretation to that model.

DACO turns the validated representation evidence into a content optimization method that uses a single guided rewrite that remains competitive with the comparison methods across the primary evaluators and domains while preserving factual content. The functionally validated components also yield practical guidance without exposing the underlying representation analysis to the user. Additional evaluations with Llama 4 Scout, Llama-3.1-405B-Instruct Grattafiori et al. (2024), Gemini 3.6 Flash, and human annotators test generalization. The closed-evaluator results place DACO within the observed comparison range; the additional human evaluation provides a supplementary check on preference and content preservation.

7 Conclusion and Future Work

Content homogenization can preserve evaluator accuracy while changing which content features predict the evaluator's decisions. This finding shows that stable accuracy can hide meaningful changes in how an evaluator responds to AI-rewritten content. The features that remain important after homogenization can also provide useful guidance for content optimization. DACO uses these identified features in a single guided rewrite while preserving the source information and remaining competitive with the comparison methods. The external evaluations extend the study beyond the primary setting, with human evaluation serving as an additional validation check. Functional validation shows that the identified features can provide practical guidance, while stronger causal evidence remains an important direction for future work. Future work can also examine more evaluator families and domains, broaden human evaluation, and further improve representation-guided content optimization.

Ethics Statement

DACO is constrained to preserve genuine source content and adds no fabricated achievements, evidence, qualifications, facts, numbers, or entities. Protected source literals are preserved mechanically and validated exactly, with zero protected-literal failures in generation. The hiring application is treated as a high-stakes evaluation setting within the broader study, not as permission to misrepresent applicant information. The Djinni data consist of anonymized CVs and job postings Drushchak & Romanyshyn (2024). The human evaluation used two annotators under a blind, counterbalanced protocol, and their judgments were combined by a fixed deterministic rule with no subjective adjudication.

Reproducibility Statement

All model, layer, PCA dimension, component, and guidance choices use training and validation data only, and the test set is held out from these choices. Every pair is evaluated in both A/B orders and remapped to the true candidate identities, which controls position bias. Protected source literals are preserved mechanically through opaque placeholders, deterministic restoration, and exact post-restoration validation. The statistical methods are defined where they are used in Section 3.1 and Section 3.2, and the setup is in Section 4. RAF runs under its published protocol, with full settings in Appendix A. Exact prompts, hardware details, model revisions, and full implementation settings are in Appendix A.

Required AI Use Statement

This paper studies LLM evaluators and uses LLM rewriting models as described in Section 4. The rewriting models gpt-oss-120b Agarwal et al. (2025) and Llama 4 Scout Meta (2025) produced the homogenized and rewritten content. The evaluators Gemma 4 12B Team et al. (2026), Ministral 3 14B Liu et al. (2026), Phi-4 14B Abdin et al. (2024), Llama-3.1-405B-Instruct Grattafiori et al. (2024), and Gemini 3.6 Flash Google (2026) produced the preference and ranking judgments analyzed in the paper. gpt-oss-120b also generated the guidance statements from high- and low-projection training examples, and each statement was retained only after independent held-out functional validation. AI was used only to correct grammar and minor writing errors in the manuscript.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Preet Baxi, Jiannan Xu, Jane Yi Jiang, and Stefanus Jasin. Prompt injection in automated résumé screening with large language models: Single and multi-injection settings. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 2942–2953, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.142. URL https://aclanthology.org/2026.findings-acl.142/.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Lee Cohen, Connie Hong, Jack Hsieh, and Judy Hanwen Shen. Two tickets are better than one: Fair and accurate hiring under strategic LLM manipulations. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of

Proceedings of Machine Learning Research, pp. 11142–11172. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/cohen25a.html.

Anil R Doshi and Oliver P Hauser. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science advances*, 10(28):eadn5290, 2024.

Nazarii Drushchak and Mariana Romanyshyn. Introducing the djinni recruitment dataset: A corpus of anonymized cvs and job postings. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP)@ LREC-COLING 2024*, pp. 8–13, 2024.

Bradley Efron. Better bootstrap confidence intervals. *Journal of the American statistical Association*, 82(397):171–185, 1987.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5988–6008. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/ethayarajh22a.html.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of machine learning research*, 20(177):1–81, 2019.

Google. Gemini 3.6 Flash — Gemini API — Google AI for Developers — ai.google.dev. https://ai.google.dev/gemini-api/docs/models/gemini-3.6-flash, 2026. [Accessed 03-09-2026].

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Andreas Köpf, Yannic Kilcher, Dimitri Von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations-democratizing large language model alignment. *Advances in neural information processing systems*, 36:47669–47681, 2023.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Aounon Kumar and Himabindu Lakkaraju. Manipulating large language models to increase product visibility. *arXiv preprint arXiv:2404.07981*, 2024.

Alexander H Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, et al. Ministral 3. *arXiv preprint arXiv:2601.08584*, 2026.

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*, pp. 2511–2522, 2023.

Meta. llama-models/models/llama4/MODEL_CARD.md at main · meta-llama/llama-models — github.com. https://github.com/meta-llama/llama-models/blob/main/models/llama4/{M}{O}{D}{E}{L}_{C}{A}{R}{D}.md, 2025. [Accessed 03-09-2026].

Jelena Mirkovic, Brian Kocoloski, and David Balenson. Enabling reproducibility through the sphere research infrastructure. *Usenix; login Magazine*, 2024.

Kibum Moon, Adam E Green, and Kostadin Kushlev. Homogenizing effect of large language models (llms) on creative diversity: An empirical comparison of human and chatgpt writing. *Computers in Human Behavior: Artificial Humans*, pp. 100207, 2025.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in llm-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 292–314, 2025.

Zhivar Sourati, Farzan Karimi-Malekabadi, Meltem Ozcan, Colin McDaniel, Alireza Ziabari, Jackson Trager, Ala N Tak, Meng Chen, Fred Morstatter, and Morteza Dehghani. The shrinking landscape of linguistic diversity in the age of large language models. *Nature Human Behaviour*, pp. 1–18, 2026.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, et al. Large language models are not fair evaluators. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9440–9450, 2024.

Yumeng Wang, Catherine Chen, and Suzan Verberne. Ranksteer: Activation steering for pointwise llm ranking. *arXiv preprint arXiv:2602.03422*, 2026.

Tiancheng Xing, Jerry Li, Yixuan Du, and Xiyang Hu. Are llms reliable rankers? rank manipulation via two-stage token optimization. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9120–9132, 2026.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, volume 2024, pp. 1651–1678, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges. In *International Conference on Learning Representations*, volume 2025, pp. 51257–51296, 2025.

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.

A Experimental Details

A.1 Data, Models, and Revisions

The study uses SHP, OASST1, and Djinni with training/validation/test splits of 360/120/120, 180/60/60, and 100/40/60, respectively. Djinni preserves the Product Manager/UI/UX balance within each split. Source selection uses deterministic source-level ordering, and non-rewriteable fields are skipped within the same dataset, split, and stratum. The test split is excluded from all model, layer, PCA dimension, component, and guidance selection.

The primary evaluators are *google/gemma-4-12B-it*, *mistralai/Ministral-3-14B-Instruct-2512*, and *microsoft/phi-4*, using revisions `707f0a3b8a3c7ad586ed01e27eafbad8a27dd0f7`, `29439f81c2be264d8d393273f99e7db9c0961120`, and `2db69c1c3e91a05d2c64a3185acfbaf36f744e25`, respectively. The local experiments use four NVIDIA L40S GPUs on the NSF SPHERE research infrastructure. Llama 4 Scout is a mixture-of-experts (MoE) model with 109 billion total parameters, 17 billion active parameters, and 16 experts Meta (2025). The supplementary cross-scale evaluation used *meta-llama/Llama-3.1-405B-Instruct* through NDIF/NNsight. Randomized procedures use fixed deterministic seeds, with 42 as the base seed.

Algorithm 1 DACO: Discriminability-Aware Content Optimization

Require: Training pairs D_{disc}, validation pairs D_{val}, evaluator E, rewriting model G, source text x
Ensure: Optimized output x_{DACO}

```
Generate the homogenized condition x_H from x under the factual-preservation constraints
Evaluate natural and homogenized pairs in both A/B orders with E
Construct order-symmetrized pair representations and evaluator margins
Fit the shared PCA basis and the natural and homogenized ridge surrogates
Compute predictive-importance changes ΔPI_k
Select the strongest supported components with preserved or increased predictive importance
for each selected component v_k do
    Rank training examples by projection r_i = z_i^⊤ v_k
    Generate guidance g_k from the highest- and lowest-projection examples
    Apply g_k to the held-out validation examples
    Compute mean projection movement Δ̄_k
    if Δ̄_k > 0 and g_k passes the guidance-validity checks then
        Retain g_k as functionally validated guidance
    end if
end for
Let g* be the validated guidance for the selected DACO component
x_DACO ← G(x_H, g*) under the same factual-preservation constraints
Evaluate x_DACO with E for preference and component projection change
return x_DACO
```

Table 5: Dataset identifiers and revisions used in the experiment.

Dataset	Repository	Revision
SHP	*stanfordnlp/SHP*	`e94b5f32602712d78ed4 94fe79105b1959396686`
OASST1	*OpenAssistant/oasst1*	`fdf72ae0827c1cda404a ff25b6603abec9e3399b`
Djinni applicant profiles	*lang-uk/recruitment-dataset-candidate-profiles-english*	`86255174c6c378a2b52c d7e81d79002c64d899a6`
Djinni job descriptions	*lang-uk/recruitment-dataset-job-descriptions-english*	`b56a6054c10f1266a141 e37c8df66c79ff2863af`

SHP pairs are stratified across the 18 dataset domains, with no SHP item reused across selected pairs. OASST1 uses English assistant responses with valid ranking information; for each selected prompt, the pair contains the highest-ranked response and the next response with a different rank, with no prompt reused across selected pairs. Djinni preserves the Product Manager/UI/UX balance. Each target applicant is evaluated with four fixed competitors in a five-applicant pool matched by position, experience group, and language. Applicant profiles are not reused across selected Djinni pools, and job descriptions are not shared across training, validation, and test splits.

A.2 Generation and Evaluation Prompts

The prompt blocks and instruction text below reproduce the executed configuration verbatim. We retain their original wording for reproducibility.

Pairwise evaluator prompt.

> You are evaluating two candidates under the same evaluation context.
> **TASK:**
> {task}
> **EVALUATION CONTEXT:**
> {context}
> **CANDIDATE A:**

{candidate_a}
END CANDIDATE
CANDIDATE B:
{candidate_b}
END CANDIDATE
Respond with exactly A or B.

For preference tasks, the task instruction is: "Select the candidate that is the better response to the user input." For hiring tasks, it is: "Select the candidate that is the better fit for the job description based only on the information provided."

Homogenization prompt.

> Rewrite the MASKED SOURCE TEXT into a clear, polished, professionally structured, and stylistically standardized form. The rewrite must preserve the substantive content exactly in meaning. Preserve every factual claim, name, entity, qualification, achievement, evidence, limitation, uncertainty, and caveat. Do not add new claims, facts, achievements, qualifications, evidence, examples, recommendations, or unsupported information. Do not remove substantive information or infer information from the context. Approximately preserve the amount of substantive content. Improve wording, organization, clarity, consistency, and presentation only.
>
> Protected literals are replaced by opaque tokens before generation. Every protected token must be preserved exactly once and in source order. The model must not create new protected tokens or numeric digits. The original values are restored deterministically after generation and checked exactly.
>
> The context is supplied only to clarify what the source refers to and may not be added to the rewrite. The model returns only the rewritten masked source text between `BEGIN REWRITE` and `END REWRITE`.

Homogenization uses gpt-oss-120b with temperature 0, low reasoning effort, one generation pass, and a maximum generation budget of 4096 tokens. Protected digits, URLs, emails, and other literal-bearing spans are restored deterministically after generation.

Guidance and comparison instructions.

Table 6: Executed generation instructions for guidance and comparison conditions.

Condition	Instruction
Guidance generation	Given HIGH- and LOW-projection examples, identify the textual property that most consistently distinguishes the groups. Return one short, actionable textual direction descriptor and do not mention PCA, vectors, activations, models, or scores.
B1	Produce a strong alternative professional rewrite while preserving the exact underlying content and facts.
B2	Improve distinctiveness for an evaluation environment where many inputs have already been strongly AI-polished and stylistically homogenized. Preserve all meaning and facts. Do not use internal representation information.
A1	Improve the text while emphasizing the validated natural-condition representation properties.
A2	Improve the text while emphasizing matched non-target representation properties.
DACO	Rewrite the input once while preserving meaning and all factual claims. Do not invent achievements, numbers, entities, qualifications, evidence, or other facts. Approximately preserve length. Improve the text specifically along the validated homogenization-resistant evaluator property.

Guidance generation uses 20 highest-projection and 20 lowest-projection training examples per component. Each generated guidance statement is tested on 12 held-out validation examples. B1

Table 7: Validation-selected representation layers and held-out surrogate performance for the three primary evaluators.

Evaluator	Transformer layer	Natural R^2	Homogenized R^2
Gemma 4 12B	28	0.6754	0.6740
Ministral 3 14B	19	0.9391	0.9406
Phi-4	22	0.8856	0.8871

uses B0 plus three additional rewrites with temperature 1.0, top-$p = 1.0$, and deterministic seeds. All four rewrites must be distinct.

A.3 Representation and Causal Analysis

The PCA dimension is selected from $K \in \{8, 16, 32, 64, 128, 256, 512, 1024, 1879\}$ using validation performance; the finalized shared PCA analysis uses $K = 512$ for all three primary evaluators. Ridge models are fit separately to natural and homogenized representations. Ridge regularization is selected from $\{0.01, 0.1, 1, 10, 100\}$ using training/validation data. Permutation importance uses 30 repeats. Paired BCa intervals use 2000 bootstrap resamples and 95% confidence. The analysis uses random seed 42.

The representation layer is selected using training and validation data by maximizing the lower of the natural- and homogenized-condition validation R^2 values; the test set is excluded from this choice. Table 7 reports the selected transformer layer and the held-out validation performance of the corresponding ridge surrogates. All three primary evaluators satisfy the predefined surrogate validity criterion.

For a selected normalized PCA component v_k and A/B order o, the rescue intervention restores the natural-to-homogenized difference in pair projection:

$$\delta^{(o)} = \left[(d_N^{(o)})^\top v_k - (d_H^{(o)})^\top v_k \right] v_k, \tag{8}$$

$$h'_{A,H} = h_{A,H} + \frac{1}{2}\delta^{(o)}, \qquad h'_{B,H} = h_{B,H} - \frac{1}{2}\delta^{(o)}. \tag{9}$$

For components with preserved or increased importance, the ablation intervention removes the homogenized pair component:

$$p^{(o)} = \left[(d_H^{(o)})^\top v_k \right] v_k, \tag{10}$$

$$h'_{A,H} = h_{A,H} - \frac{1}{2}p^{(o)}, \qquad h'_{B,H} = h_{B,H} + \frac{1}{2}p^{(o)}. \tag{11}$$

Each selected-component intervention is compared with a same-size matched random orthogonal intervention. Causal effects use paired BCa intervals with 2000 resamples and BH-FDR correction at $q < 0.05$.

A.4 RAF, Closed-Model, and Human Protocols

RAF uses the reference implementation at revision `bfeb12bddd185d588da677b1230e7f90c347adce`. It is evaluated only on the Djinni test set with Phi-4. The evaluation contains 60 target applicants, a five-applicant pool consisting of the target applicant plus four fixed competitors, 10 optimization steps per target, top-$k = 64$, a maximum optimized prompt length of 10 tokens, and base seed 42 with deterministic per-target seeds. RAF's mean rank was 2.7333 and median rank 3.0000 in the five-applicant pool.

The closed black-box evaluation uses Gemini 3.6 Flash through the asynchronous Batch API with `thinkingLevel=minimal`, temperature 0, and `maxOutputTokens=16`. All six conditions are evaluated in both A/B and B/A orders. The evaluation contained 5040 requests corresponding to 2520 paired metrics, or 420 paired evaluations per method. Selection rate and AB/BA order consistency are computed after candidate-identity remapping.

Table 8: Llama 4 Scout results on the completed subset.

Metric	Gemma 4 12B	Ministral 3 14B	Phi-4
Preference accuracy, natural	54.02%	57.47%	63.22%
Preference accuracy, Scout-homogenized	55.17%	57.47%	65.52%
Mean signed margin, natural	2.4278	1.7410	3.0489
Mean signed margin, Scout-homogenized	1.7721	1.4781	2.4282
Djinni pairwise win rate, natural	34.62%	41.35%	31.73%
Djinni pairwise win rate, Scout-homogenized	37.50%	38.46%	36.54%

Table 9: PCA component classification across evaluators. Each evaluator contains 512 shared-PCA components.

Evaluator	Decreased	Increased	Preserved	Unsupported
Gemma 4 12B	29	41	40	402
Ministral 3 14B	40	28	68	376
Phi-4	29	26	50	407

The human evaluation contains 180 samples, with 60 each from SHP, OASST1, and Djinni. Two independent annotators evaluate every sample blindly. B0 and DACO are presented as Candidate A and Candidate B. Placement is balanced at 90 B0-as-A and 90 DACO-as-A overall, including 30/30 within each dataset, and the second annotator receives the opposite A/B order. The preference question is "Which version better satisfies the evaluation context while preserving the underlying source content?" The preservation question is "Does the transformed version preserve the factual content of the source without adding unsupported facts?" Preference labels are A better, Tie, and B better; preservation labels are Preserved and Not Preserved.

Agreement is calculated before aggregation. Matching preference judgments are retained. A directional judgment paired with Tie retains the directional judgment. Opposing A/B judgments remain unresolved. Preservation disagreements remain unresolved. No third adjudicator is used.

B Additional Results

The evaluation with the second rewriting model, Llama 4 Scout, targeted 120 items: 60 SHP, 30 OASST1, and 30 Djinni. Generation completed for 113 items, including 58 SHP, 29 OASST1, and 26 Djinni items; no replacement or imputation was used. The completed SHP and OASST1 subset contains 87 preference pairs. The completed Djinni subset contains 26 target applicants evaluated against four fixed competitors, giving 104 pairwise comparisons per condition. These completed items produce 382 evaluation instances. Table 8 reports the results for the three primary evaluators.

The supplementary Llama-3.1-405B-Instruct analysis used 60 SHP, 30 OASST1, and 30 Djinni items at `model.layers.63`. A deterministic source-level split assigned 40 items to training (20 SHP, 10 OASST1, 10 Djinni), 20 to validation (10, 5, 5), and 60 to test (30, 15, 15). The shared PCA analysis used 128 components. On the held-out SHP and OASST1 preference subset ($n = 45$), preference accuracy was 68.89% in the natural condition and 57.78% after homogenization. The corresponding mean signed margins were 3.3653 and 2.1583, with a paired change of -1.2069 and a 95% BCa interval of $[-2.0269, -0.6010]$. Across the 105 held-out test comparisons, the median representation-norm ratio was 1.0455 with a 95% BCa interval of $[1.0118, 1.1694]$. The held-out surrogate had validation $R^2 = -0.2080$ for the natural condition and -0.0329 for the homogenized condition. The surrogate therefore did not satisfy the predefined validity criterion used for component interpretation. Accordingly, the Llama-3.1-405B-Instruct branch did not proceed to causal intervention, and 0 intervention jobs were run.

None of the 12 component-specific effects reaches the predefined BH-FDR threshold of $q < 0.05$.

Table 11 complements Table 2: the main table reports paired signed-margin differences with uncertainty, whereas this table reports absolute preference accuracies. DACO's corresponding mean signed margins are 3.2455, 2.0611, and 3.0181 for Gemma, Ministral, and Phi-4, respectively.

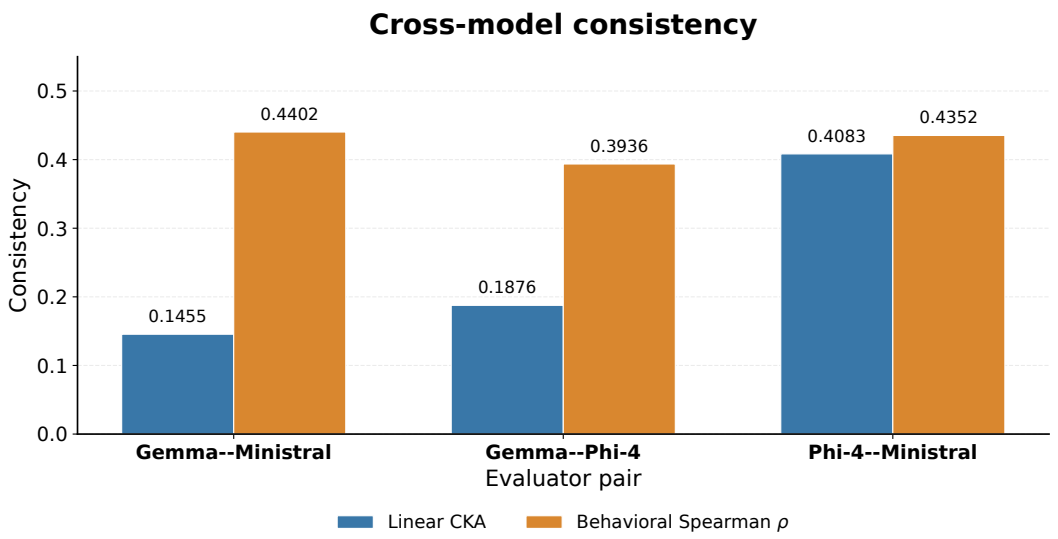

Figure 2: Cross-model consistency of the homogenization-induced shift. Linear CKA measures representation-shift similarity, and Spearman ρ measures behavioral agreement. Behavioral correlations use $n = 180$ paired examples per evaluator pair.

Table 10: Causal validation of the selected PCA components. Effect is the mean evaluator-margin change for the selected component minus its matched random orthogonal control. Each test uses $n = 420$ paired examples.

Evaluator	Component	Class	Intervention	Effect [95% BCa CI]	BH-FDR q
Gemma	PC0	Decreased	Rescue	-0.0026 $[-0.0149, 0.0098]$	0.716
Gemma	PC42	Decreased	Rescue	$+0.0036$ $[-0.0071, 0.0132]$	0.714
Gemma	PC13	Preserved	Ablation	$+0.0030$ $[-0.0077, 0.0130]$	0.714
Gemma	PC12	Preserved	Ablation	$+0.0085$ $[-0.0043, 0.0229]$	0.446
Ministral	PC26	Decreased	Rescue	$+0.0087$ $[-0.0023, 0.0195]$	0.446
Ministral	PC93	Decreased	Rescue	-0.0028 $[-0.0132, 0.0081]$	0.714
Ministral	PC0	Preserved	Ablation	-0.0006 $[-0.0138, 0.0141]$	0.959
Ministral	PC2	Increased	Ablation	-0.0036 $[-0.0165, 0.0080]$	0.714
Phi-4	PC36	Decreased	Rescue	$+0.0037$ $[-0.0027, 0.0095]$	0.446
Phi-4	PC44	Decreased	Rescue	-0.0046 $[-0.0109, 0.0021]$	0.446
Phi-4	PC0	Preserved	Ablation	-0.0374 $[-0.0709, -0.0071]$	0.252
Phi-4	PC3	Preserved	Ablation	$+0.0138$ $[-0.0019, 0.0278]$	0.336

C Practical Guidance and Human Materials

The analysis of practical guidance begins with the 96 Ministral components classified as preserved or increased. For each component, gpt-oss-120b receives the 20 highest-projection and 20 lowest-projection training examples and produces one short actionable guidance statement. Each statement is then tested on 12 held-out validation examples from the representation-analysis validation split. A statement is retained when its mean projection movement is strictly positive and it satisfies the plain-language guidance constraints. This procedure produces 34 validated guidance statements. Only those 34 guidance statements are then supplied to gpt-oss-120b for conservative semantic grouping; PCA identifiers, ranks, scores, labels, and other metadata are excluded. Statements are grouped only when they express materially the same actionable advice. This produces the five practical insights reported in Table 3. The complete validated-guidance records and grouping evidence are supplied with the machine-readable supplementary results.

The human evaluation uses the same 180 samples described in Appendix A. Annotators evaluate preference and preservation independently before any aggregation. Raw preference agreement is 30.56% and linearly weighted Cohen's kappa is 0.0197. Preservation agreement is 99.44% for

Table 11: Absolute open-evaluator preference accuracy across methods.

Method	Gemma 4 12B	Ministral 3 14B	Phi-4
A1	60.00%	63.89%	69.44%
A2	62.78%	60.56%	62.22%
B0	62.22%	66.67%	66.11%
B1	62.78%	63.33%	65.00%
B2	60.56%	65.56%	66.11%
DACO	62.78%	63.89%	66.11%

Table 12: Djinni pairwise win rates. RAF uses its dedicated paired Phi-4/Djinni protocol.

Method	Gemma 4 12B	Ministral 3 14B	Phi-4
A1	48.33%	50.42%	46.67%
A2	50.00%	51.67%	47.50%
B0	46.25%	49.17%	45.83%
B1	50.42%	52.92%	47.50%
B2	48.75%	50.83%	48.75%
DACO	47.50%	49.17%	47.50%
RAF*	—	—	56.67%

Candidate A and 99.44% for Candidate B; Cohen's kappa is 0.0000 for both candidate positions, with one unresolved preservation disagreement for each. Deterministic aggregation gives 68 B0 selections, 58 DACO selections, 53 ties, and one unresolved preference. The original annotator judgments are retained unchanged in the supplementary human-evaluation material.