Online Resource 1: Supplementary Appendices

Debiasing LLM Recommendations via Decision-Aware Counterfactual Alignment

Information Technology and Management

This document contains Appendices A–C. References to Sections 1–6, Tables 1–7, and results or equations without an appendix prefix refer to the main article. Tables 8–11 retain their original numbering. Bibliographic reference numbers are local to this supplement.

A Proofs and Technical Details

This appendix proves the results of Section 4.2 in the order in which they appear there, and states two results that the main text uses informally: a margin condition for ranked outputs (Lemma A.1) and a characterization of the utility floor (Lemma A.2). Throughout, the policy and the fitted reward model are fixed, and expectations are over fresh policy randomness. The output encoding is bounded, so every moment used below exists. Replicate pairs are independent and identically distributed. The two responses within a pair may be dependent, provided each follows its own policy distribution.

A.1 Matched-Prompt Estimation

The mean result is linearity of expectation, the squared-effect result removes simulation variance from a squared average, and the precision result is Hoeffding's inequality combined with a union bound.

Proof of Theorem 4.1. Fix x_R and (a, a'), write $\boldsymbol{\tau} = \boldsymbol{\tau}_\theta(x_R; a, a')$, and let $\sigma_\theta^2 = \operatorname{tr}\operatorname{Cov}(\widehat{\boldsymbol{\tau}}^{(1)}) = \mathbb{E}\|\widehat{\boldsymbol{\tau}}^{(1)} - \boldsymbol{\tau}\|_2^2$.

(i) Because each response follows its policy distribution, $\mathbb{E}\widehat{\boldsymbol{\tau}}^{(b)} = \bar{\boldsymbol{\pi}}_\theta(x_R, a) - \bar{\boldsymbol{\pi}}_\theta(x_R, a') = \boldsymbol{\tau}$ regardless of the dependence within a pair, so $\mathbb{E}\widehat{\boldsymbol{\tau}}_B = \boldsymbol{\tau}$; the strong law of large numbers, applied to each coordinate, gives $\widehat{\boldsymbol{\tau}}_B \to \boldsymbol{\tau}$ almost surely.

(ii) Independence across replicates gives $\operatorname{Cov}(\widehat{\boldsymbol{\tau}}_B) = \operatorname{Cov}(\widehat{\boldsymbol{\tau}}^{(1)})/B$, so

$$\mathbb{E}\|\widehat{\boldsymbol{\tau}}_B\|_2^2 = \|\boldsymbol{\tau}\|_2^2 + \operatorname{tr}\operatorname{Cov}(\widehat{\boldsymbol{\tau}}_B) = \|\boldsymbol{\tau}\|_2^2 + \sigma_\theta^2/B.$$

The sample covariance of independent and identically distributed vectors is unbiased, and $\widehat{\sigma}_B^2$ is its trace, so $\mathbb{E}\widehat{\sigma}_B^2 = \sigma_\theta^2$. Subtracting gives $\mathbb{E}\widehat{f}_B = \|\boldsymbol{\tau}\|_2^2$.

(iii) Each coordinate $\widehat{\tau}_{B,j}$ averages B independent terms lying in an interval of width at most w. For $w > 0$, Hoeffding's inequality [1] and a union bound over the K coordinates give

$$\Pr(\|\widehat{\boldsymbol{\tau}}_B - \boldsymbol{\tau}\|_\infty \geq t) \leq 2K \exp(-2Bt^2/w^2), \qquad t > 0,$$

and setting the right side equal to α yields Equation (11). No independence across coordinates is needed, and if $w = 0$ the error is zero. □

An equivalent form, and practical use. Because $(B-1)\widehat{\sigma}_B^2 = \sum_b \|\widehat{\boldsymbol{\tau}}^{(b)}\|_2^2 - B\|\widehat{\boldsymbol{\tau}}_B\|_2^2$, expanding $\|\sum_b \widehat{\boldsymbol{\tau}}^{(b)}\|_2^2$ gives

$$\widehat{f}_B = \frac{1}{B(B-1)} \sum_{b \neq b'} \widehat{\boldsymbol{\tau}}^{(b)\top} \widehat{\boldsymbol{\tau}}^{(b')}, \tag{A.1}$$

an average of inner products between distinct replicates. By independence, each term has expectation $\mathbb{E}(\widehat{\boldsymbol{\tau}}^{(b)})^\top \mathbb{E}(\widehat{\boldsymbol{\tau}}^{(b')}) = \|\boldsymbol{\tau}\|_2^2$, which gives a second proof of part (ii). Equation (10) computes the same quantity without forming all $B(B-1)$ products. Three considerations guide its use. First, $\widehat{f}_B$ is unbiased but can be negative in a finite sample, and truncating it at zero introduces bias. Second, the nonnegative plug-in $\|\widehat{\boldsymbol{\tau}}_B\|_2^2$ is a convenient training penalty, but its excess σ_θ^2/B generally depends on θ, so penalizing it also favors policies with less generation variance. Third, averaging $\widehat{f}_B$ over profiles and contrasts drawn from P_{audit} estimates $\mathcal{F}(\theta)$, whereas averaging over a fixed, weighted panel estimates that panel's risk. The width w refers to a coordinate of the difference $\widehat{\boldsymbol{\tau}}^{(b)}$: if each arm's coordinate lies in a common interval of width w_0, then $w = 2w_0$ is valid. A within-pair coupling such as a shared seed can reduce variance without changing the target, although an arbitrary coupling need not.

A.2 Restricted-Score Stability

Proof of Lemma 4.2. For a fixed profile and candidate, let $\boldsymbol{z} = \boldsymbol{Z}_R(X_R, a, y)$ and $\boldsymbol{z}' = \boldsymbol{Z}_R(X_R, a', y)$. The restricted reward reads only $\boldsymbol{Z}_R$, and r_ψ is L_R-Lipschitz, so

$$\left|\widetilde{R}_{\phi,\psi}(X_R, a, y) - \widetilde{R}_{\phi,\psi}(X_R, a', y)\right|^2 = |r_\psi(\boldsymbol{z}) - r_\psi(\boldsymbol{z}')|^2 \leq L_R^2 \|\boldsymbol{z} - \boldsymbol{z}'\|_2^2.$$

Taking expectations under the P_{pair} that defines $\mathcal{L}_{\text{sep}}$ gives Equation (12). □

Zero separation loss therefore gives equal scores almost surely under P_{pair}. A small positive loss bounds score changes on average but not uniformly. Lemma A.1 below therefore assumes a uniform bound separately.

A.3 The End-to-End Counterfactual Bound

The proof follows the sequence in Table 3. For that path to be well defined, each hybrid rule must accept every pool it receives, and the joint construction must preserve the endpoint laws, $\boldsymbol{Y}_0 \sim P_a$ and $\boldsymbol{Y}_3 \sim P_{a'}$. Shared selection randomness η implements each endpoint's conditional selection law.

Proof of Theorem 4.3. For any coupling $(\boldsymbol{Y}, \boldsymbol{Y}')$ of P_a and $P_{a'}$, $\boldsymbol{\tau}_\theta(x_R; a, a') = \mathbb{E}(\boldsymbol{Y} - \boldsymbol{Y}')$, so Jensen's inequality gives $\|\boldsymbol{\tau}_\theta(x_R; a, a')\|_2 \le \mathbb{E}\|\boldsymbol{Y} - \boldsymbol{Y}'\|_2$; the infimum over couplings is $W_1(P_a, P_{a'})$. The hybrid endpoints $(\boldsymbol{Y}_0, \boldsymbol{Y}_3)$ form one such coupling, and the triangle inequality along the path gives

$$W_1(P_a, P_{a'}) \le \mathbb{E}\|\boldsymbol{Y}_0 - \boldsymbol{Y}_3\|_2 \le \sum_{j=1}^{3} \mathbb{E}\|\boldsymbol{Y}_{j-1} - \boldsymbol{Y}_j\|_2 = \Delta_{\text{score}} + \Delta_{\text{select}} + \Delta_{\text{pool}}.$$

This is Equation (13); squaring the bound on $\|\boldsymbol{\tau}_\theta\|_2$ and averaging over P_{audit} gives the bound on $\mathcal{F}(\theta)$. □

The stage discrepancies depend on the coupling and on the order of the swaps, so their sum is an upper bound rather than an attribution of the observed effect. The next lemma makes precise the ranked-output argument of Section 4.2. It concerns the scoring step $\boldsymbol{Y}_0 \to \boldsymbol{Y}_1$, in which the pool and the selector are held fixed and only the scores change.

Lemma A.1 (Margin condition for top-k selection)**.** *Let selection be deterministic top-k on a common pool $C = \{c_1, \dots, c_m\}$ with $1 \le k < m$, indexed in decreasing order of $\boldsymbol{s}_a$, and suppose $\|\boldsymbol{s}_a - \boldsymbol{s}_{a'}\|_\infty \le \kappa$. If $s_a(c_k) - s_a(c_{k+1}) > 2\kappa$, the selected set is the same under $\boldsymbol{s}_a$ and $\boldsymbol{s}_{a'}$. If also $s_a(c_j) - s_a(c_{j+1}) > 2\kappa$ for every $j < k$, the ordered list is the same; hence $\Delta_{\text{score}} = 0$ for a rank encoding whenever these margins hold almost surely.*

Proof. If u is ranked above v with $s_a(u) - s_a(v) > 2\kappa$, the perturbation bound gives $s_{a'}(u) - s_{a'}(v) \ge s_a(u) - s_a(v) - 2\kappa > 0$, so u stays above v. Every selected u and excluded v are separated by at least $s_a(c_k) - s_a(c_{k+1})$, which preserves membership; applying the same argument to each adjacent selected pair preserves the order. □

The boundary margin alone does not preserve order: the score vectors $(2, 1.9, 0)$ and $(1.9, 2, 0)$ differ uniformly by 0.1 and have a boundary gap of $1.9 > 0.2$, yet they select the same top-two set in opposite orders. When $k = m$, membership is unchanged trivially and only the order conditions apply.

A.4 Counterfactual Closure and the Utility Floor

The corollary requires the same realized candidate pool under the coupling, not merely a common catalog of possible options.

Proof of Corollary 4.4. Sufficiency. On a common realized pool C, equal scores give $\boldsymbol{s}_a(C) = \boldsymbol{s}_{a'}(C)$, and a cue-independent selector with shared randomness then gives $\boldsymbol{Y}_0 = \boldsymbol{Y}_1 = \boldsymbol{Y}_2 = \boldsymbol{Y}_3$ almost surely. Hence $P_a = P_{a'}$ and $\boldsymbol{\tau}_\theta(x_R; a, a') = 0$, and because the conditions hold for P_{audit}-almost every profile and contrast, $\mathcal{F}(\theta) = 0$.

Nonredundancy. Take two distinctly encoded options u and v and a rule that selects one. Each construction below satisfies two of the three conditions but violates alignment. Without common support, the pools $\{u\}$ and $\{v\}$ force different selections under a common score function and argmax

selector. Without equal scores, the common pool $\{u, v\}$ with score vectors $(1, 0)$ and $(0, 1)$ leads a common argmax to select u and v. Without a cue-independent selector, the common pool with tied scores and a tie-break favoring u under a and v under a' selects different options. □

The conditions are jointly sufficient but need not hold separately in every invariant policy, since effects at different stages can cancel. The corollary guarantees invariance without relying on such cancellation. Shared randomness aligns paired realizations, whereas independent draws from identical output distributions can still differ.

The utility floor of Section 4.2 follows from the definition of U_{flat}. The second part of the next lemma justifies computing U_{flat} from the best fixed list.

Lemma A.2 (Utility floor)**.** *If $u_\star > U_{\text{flat}}$, no $\pi \in \Pi_{\text{flat}}$ satisfies $U(\pi) \geq u_\star$. Suppose further that $U(\pi) = \mathbb{E}\, u(X_R, \boldsymbol{Y}_\pi)$ for a per-recommendation utility u, and that X_R is independent of X_A under the evaluation distribution, as in a matched design that crosses every profile with every cue. Then $U_{\text{flat}} = \sup_{\boldsymbol{y}} \bar{u}(\boldsymbol{y})$, where $\bar{u}(\boldsymbol{y}) = \mathbb{E}\, u(X_R, \boldsymbol{y})$ is the utility of always recommending $\boldsymbol{y}$.*

Proof. The first claim holds because $U(\pi) \leq U_{\text{flat}} < u_\star$ for every $\pi \in \Pi_{\text{flat}}$. For the second, a policy in Π_{flat} draws $\boldsymbol{Y}$ from a distribution Q_{X_A} that may depend on the cue but not on X_R, so independence of X_R and X_A gives $U(\pi) = \mathbb{E}_{X_A} \mathbb{E}_{\boldsymbol{Y} \sim Q_{X_A}} \bar{u}(\boldsymbol{Y}) \leq \sup_{\boldsymbol{y}} \bar{u}(\boldsymbol{y})$. The constant policy that always returns $\boldsymbol{y}$ belongs to Π_{flat} and attains $\bar{u}(\boldsymbol{y})$. □

With $u_\star = U(\pi_0) - \delta$, the first part is the relative condition of Section 4.2. The supremum is a maximum when, as in Section 5, recommendations come from a finite catalog. Ignoring X_R does not by itself imply cue invariance, since a policy in Π_{flat} may still respond to X_A.

B Supporting Empirical Evidence

This appendix documents additional evidence behind Section 5: the design of the motivating audit (Appendix B.1), precise definitions of every outcome (Appendix B.2), the absolute profile responses behind the response index (Appendix B.3), and the complete signed STEM contrasts (Appendix B.4).

B.1 Design of the Motivating Audit

The audit applies the college-major task to the same five locally served models before any mitigation. It comprises 4,250 recommendations from the 100-major catalog and covers four attribute dimensions: gender identity, sexual orientation, socioeconomic status, and race or ethnicity. Matched prompts keep the academic information fixed and change only the demographic description. The audit profiles are separate from the 20 families used to develop and evaluate DACA, and the audit is distinct from Zheng's GPT-3.5-Turbo study [2].

For each model, the audit records three summaries of matched lists, defined in Table 8: the mean number of shared majors, movement, and the absolute STEM shift. These appear in Table 7

and are discussed in Section 5.3. The audit also tests for a consistent directional STEM effect, combining paired uncertainty estimates, a prespecified practical effect threshold, and false-discovery-rate adjustment for multiple comparisons [3]. A test is flagged only if it meets this combined rule. No test was flagged for any of the five models. The audit therefore separates two findings: every model changes which majors it recommends when only the demographic description changes, but none shows a consistent signed STEM penalty. This distinction motivates a target defined on the complete list rather than on an exposure summary.

B.2 Outcome Definitions and Reporting Scales

Table 8 defines every outcome reported in Section 5 and Appendix C, with its scale, preferred direction, and where it appears. The results use three reporting conventions.

First, several statistics are multiplied by 100 for readability, which changes the scale but not the meaning. An academic fit of 86.4, for example, is a normalized discounted cumulative gain of 0.864 against the rubric's ideal ranking, not 86.4% correct advice.

Second, shared majors are reported as counts rather than converted from an overlap coefficient. For two ten-major lists A and B, the Jaccard coefficient is $J = |A \cap B|/|A \cup B| = |A \cap B|/(20 - |A \cap B|)$, so each pair's shared count is $20J/(1 + J)$. Because this transformation is nonlinear, the mean shared count is not the transform of the mean Jaccard coefficient. The tables therefore report counts directly.

Third, movement and exposure discrepancy summarize realized matched outputs. They are related to, but are not estimators of, the distributional distance $W_1(P_a, P_{a'})$ and the audit risk $\mathcal{F}(\theta)$ of Sections 3–4, whose replicate-based estimators are those of Theorem 4.1.

B.3 Absolute Profile Response

The comparison tables report profile response as an index relative to original generation. Table 9 gives the absolute statistic P behind that index for the common-policy comparison of Table 5. An index above 100 means that a policy differentiates among academic profiles more than original generation does. For example, conventional reward's index of 101.2 reflects $P = 0.333$ against $P_0 = 0.329$. The index is computed from unrounded values of P, so a ratio of the rounded entries can differ in the last digit, as for cue omission, where 0.301/0.329 gives 91.5 against the reported 91.6.

B.4 Signed STEM Differences by Cue Pair

Table 10 reports, for each of the ten contrasts in $\mathcal{C}_A$, the first cue's rank-weighted STEM exposure minus the second cue's, averaged equally across models on the held-out panel. Positive entries indicate greater STEM exposure under the first cue. Only original generation and conventional reward are shown, because cue omission, aligned reward, and robust DACA are zero in every row. For omission and robust DACA this follows from their construction, since matched variants receive

Table 8: Outcome definitions. "Better" gives the preferred direction: profile response is to be preserved rather than maximized, and source distance describes the extent of an intervention rather than its quality.

Outcome	Scale	Better	Definition	Reported in
Movement	$100M$	Lower	Transport discrepancy between matched ranked lists; zero only when they contain the same majors in the same order	Tables 5–7; acceptance (mean and worst case)
Shared majors	Count out of 10	Higher	Mean number of majors common to two matched ten-major lists	Table 7
STEM gap (audit: STEM shift)	Percentage points	Lower	Absolute difference in rank-weighted STEM exposure between matched lists; signed values in Appendix B.4	Tables 5–7
Exposure discrepancy	$100E$	Lower	Euclidean discrepancy between matched rank-exposure vectors over the 100 majors	Table 5; acceptance (exposure shift)
Academic fit	$100U$	Higher	Normalized discounted cumulative gain against the academic-fit rubric [4]; 100 is the rubric's ideal ranking	Tables 5 and 7; acceptance
Lowest-cue fit	$100U_{\min}$	Higher	Academic fit of the least favorably treated cue variant	Table 5; acceptance
Profile response	Index $100P/P_0$; absolute P	Preserve	Differentiation among legitimately different academic profiles; P_0 is the value for original generation	Tables 5 and 9; acceptance (P)
Source distance	$100T$	–	Revision of the original recommendations	Table 5; acceptance (ceiling)
Candidate coverage	$100C$	Higher	Coverage of the candidate options	Acceptance only

Table 9: Absolute profile response on held-out profiles for the common-policy comparison, averaged equally across models, and the resulting index reported in Table 5.

Policy	Absolute response P	Index $100P/P_0$
Original generation	0.329	100.0
Cue omission	0.301	91.6
Conventional reward	0.333	101.2
Aligned reward	0.332	100.9
Robust DACA	0.329	100.0

the same output. The values are descriptive and carry no pair-specific significance tests. The motivating audit's adjusted tests concern a different sample. Section 5.5 discusses the patterns, whose signs change with the form of the cue.

Table 10: Signed STEM-exposure differences on held-out profiles, in percentage points, grouped by comparison type. Positive entries indicate greater exposure under the first cue. Cue omission, aligned reward, and robust DACA are zero in every row.

Cue contrast (first minus second)	Original generation	Conventional reward
Gender		
Woman − man (explicit)	−1.45	−0.02
Woman − man (pronouns)	+0.07	0.00
Nonbinary − man (explicit)	−2.95	−0.07
Nonbinary − man (pronouns)	+0.77	0.00
Race		
Asian − White (explicit)	+0.02	0.00
Asian − White (paraphrase)	+0.93	0.00
Black − White (explicit)	−1.41	+0.05
Black − White (paraphrase)	−2.30	0.00
Race within gender		
Asian woman − Black woman	−0.59	+0.05
Asian man − Black man	+0.91	0.00

C Recommendation-Policy Selection and Held-Out Testing

A *recommendation policy* here is the complete rule used with a fixed base model: the prompt treatment, candidate-pool construction, fitted scorer, and final selection rule. For each model, the five alternatives of Table 5 were compared on validation profiles against the prespecified requirements of Table 11, and the feasible alternative with the highest external utility was selected. When none was feasible, the original policy was retained as an uncertified fallback. The selected rules, and the reasons for the two that are not robust DACA, appear in Table 7 and Section 5.6. All choices were frozen before any held-out profile was examined.

C.1 Prespecified Requirements and Held-Out Results

Table 11 lists every requirement together with the estimate and simultaneous confidence bound for each model's selected rule. Upper limits apply to upper bounds and lower limits to lower bounds. A rule is accepted only if every requirement passes. Profile response is tested on its absolute scale P rather than as an index. The bounds are simultaneous across the application's tests and differ from the fixed-profile precision bound of Theorem 4.1. The zero cue discrepancies of the omission and robust rules follow from their common-output construction (Corollary 4.4), so for these rules the informative entries are fit, responsiveness, source distance, and coverage. Two kinds of shortfall should be distinguished. DeepSeek's omission rule and Qwen's fallback violate requirements on

their point estimates, whereas the robust rules for GPT-OSS and Gemma meet the source-distance ceiling on point estimates but not on upper bounds. Their compliance is therefore unestablished rather than refuted.

Table 11: Held-out acceptance tests for the rules selected on validation profiles, with models ordered as in Table 7. Entries are estimates with simultaneous bounds in brackets, in unscaled units (multiply by 100 for the score points of Section 5); UB denotes an upper bound and LB a lower bound. Bold entries do not satisfy the requirement.

Criterion	Required	DeepSeek	Qwen	GPT-OSS	Llama	Gemma
Selected rule		Omission	Original	Robust	Robust	Robust
Mean movement (UB)	≤ 0.03	0.000 [0.000]	**0.143 [0.240]**	0.000 [0.000]	0.000 [0.000]	0.000 [0.000]
Worst movement (UB)	≤ 0.08	0.000 [0.000]	**0.586 [0.586]**	0.000 [0.000]	0.000 [0.000]	0.000 [0.000]
Exposure shift (UB)	≤ 0.03	0.000 [0.000]	**0.140 [0.219]**	0.000 [0.000]	0.000 [0.000]	0.000 [0.000]
Mean fit (LB)	≥ 0.60	**0.532 [0.420]**	**0.663 [0.568]**	0.800 [0.670]	0.831 [0.770]	0.900 [0.838]
Lowest-cue fit (LB)	≥ 0.55	**0.532 [0.377]**	**0.640 [0.512]**	0.800 [0.618]	0.831 [0.751]	0.900 [0.814]
Profile response (LB)	≥ 0.05	0.339 [0.167]	0.269 [0.076]	0.353 [0.113]	0.422 [0.155]	0.348 [0.140]
Source distance (UB)	≤ 0.50	0.319 [0.411]	0.000 [0.000]	**0.455 [0.538]**	0.371 [0.450]	**0.473 [0.553]**
Candidate coverage (LB)	≥ 0.15	0.264 [0.223]	0.604 [0.473]	0.609 [0.516]	0.623 [0.511]	0.581 [0.511]
All requirements met		No	No	No	Yes	No

References

[1] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, 1963. https://doi.org/10.1080/01621459.1963.10500830.

[2] A. Zheng, "Dissecting bias of ChatGPT in college major recommendations," *Information Technology and Management*, vol. 26, no. 4, pp. 625–636, 2025. https://doi.org/10.1007/s10799-024-00430-5.

[3] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B*, vol. 57, no. 1, pp. 289–300, 1995. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

[4] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Transactions on Information Systems*, vol. 20, no. 4, pp. 422–446, 2002. https://doi.org/10.1145/582415.582418.