Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Debiasing LLM Recommendations via Decision-Aware Counterfactual Alignment

Alex Zheng

alexzhen@andrew.cmu.edu

Carnegie Mellon University

Yuhan Zheng

University of Nottingham Ningbo

Research Article

Keywords: large language models, personalized recommendations, bias mitigation, counterfactual alignment, decision support

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11505733/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Debiasing LLM Recommendations via Decision-Aware Counterfactual Alignment

Abstract

Large language models (LLMs) increasingly recommend options for consequential decisions, but information irrelevant to a decision can change which options they present and how they rank them. Mitigating this bias requires removing that influence without suppressing legitimate personalization. We develop Decision-Aware Counterfactual Alignment (DACA), which makes the recommendation delivered to the user the object of mitigation. DACA compares versions of decision cases that differ only in an irrelevant cue and uses the same candidate options for each. It trains a reward model on task-relevant evidence rather than the irrelevant cue, directly penalizes cue-induced differences in the final recommendation, and accepts an update only if it meets an independently measured utility requirement. We derive an end-to-end bound linking cue effects in generation, scoring, and selection to differences in final recommendations. We show that cue invariance at all three stages guarantees exact alignment, while invariance at any two does not. In a controlled college-major experiment with five open-weight LLMs, changing only a race or gender description replaces about 2.5 of ten recommended majors. DACA provides otherwise identical academic profiles the same majors in the same order, raises average rubric-based academic fit from 63.5 to 86.4 out of 100, and retains responsiveness to academic differences. Omitting the cue from the prompt also equalizes lists, but compromises fit and reduces responsiveness. Ablations show that reward learning supplies most of the fit gain, while a common candidate pool and final-output alignment remove changes in named majors and ranks that persist after the STEM-exposure gap closes.

Keywords: large language models; personalized recommendations; bias mitigation; counterfactual alignment; decision support

1 Introduction

Large language models (LLMs) increasingly support consequential choices, including educational and career guidance, financial advice, and clinical decision support [1, 2, 3, 4]. In these roles, they summarize information and determine which alternatives a person sees and in what order. Even when

the person retains final authority, the recommendation shapes the opportunities considered. If an otherwise unchanged profile receives different recommendations because of an irrelevant demographic description, the model has altered the decision for reasons unrelated to the evidence that should govern it. The practical concern extends beyond biased language or an unequal summary statistic, such as a disparity score. It also concerns differences in the opportunities presented to the same decision case.

This problem has precedent in algorithmic decision support. A widely used health-management algorithm understated the needs of Black patients relative to similarly ill White patients because it used healthcare spending as a proxy for medical need [5]. A gender-neutral advertising campaign for STEM careers reached fewer women because cost-optimized delivery favored audiences that were cheaper to reach [6]. In both cases, an apparently neutral objective produced unequal allocations because the signal being optimized did not represent the purpose of the decision. For an LLM recommender, defining recommendation quality is therefore part of addressing bias, not a separate performance concern.

LLMs add further routes by which demographic information can reach an output. Representations learned from text may encode social associations [7], and identity can be conveyed through names, dialect, or context rather than an explicit demographic field. Prior studies document race-sensitive clinical advice [1], dialect-dependent occupational and criminal judgments [8], and name-dependent employment and salary recommendations [2]; other evidence reveals religious framing in financial advice and institutional prestige in candidate evaluation [4, 9]. Together, these findings indicate that mitigation must account both for the forms in which a cue is expressed and for the recommendations it changes.

We define *LLM decision bias* as an unauthorized effect of an *audited cue* on the final recommendation when decision-relevant information is held fixed. The audited cue is the attribute or textual expression under test, such as a demographic label, a name, or a pronoun. A *decision standard*, fixed in advance, specifies which information may legitimately influence the output and which cue effects are prohibited. It may derive from law, institutional policy, professional norms, or the purpose of the decision. Under an academic-fit standard, for example, a student's preparation, interests, and career goals should shape the recommended majors, whereas an otherwise irrelevant demographic description should not. The target is thus *within-profile consistency*, not uniform advice: the same decision case should receive the same recommendation whatever its cue, while cases that differ in legitimate information may receive different recommendations.

Four challenges arise to achieve this target. First, simple comparisons between groups (e.g., of certain demographics) can mix cue effects with legitimate differences between profiles. Second, deleting an explicit attribute can leave related information elsewhere in the text [8, 2]. Third, aggregate exposure measures can conceal substitutions: e.g., two lists can contain the same share of STEM majors yet recommend different majors or rank them differently. Fourth, and most consequential for mitigation, the LLM itself generates the options that later scoring and ranking rules receive. Fair-ranking methods, for example, optimize exposure over a given candidate set [10],

and no reranker can restore a suitable option that was never generated. Bias can therefore enter at three stages: which candidates are generated, how they are scored, and which are selected and in what order. Effective mitigation must govern all three while maintaining the quality of what is recommended.

We introduce Decision-Aware Counterfactual Alignment (DACA), which makes the delivered recommendation the object of mitigation. DACA compares two versions of the same decision case that differ only in a prespecified audited cue, following the controlled-comparison logic of correspondence audits [11]. Repeated generations under each version separate a persistent cue effect from ordinary sampling variation. The matched cases then govern every stage of learning, so that generation, scoring, and selection are held to one definition of permissible variation. Two learned components carry this out. A *reward model* estimates how well a candidate fits the legitimate profile. The *recommendation policy* comprises the complete rule for generating, scoring, and selecting the delivered output and is trained against that reward. DACA restricts the reward model's use of the audited cue and directly penalizes cue-induced differences in the policy's final recommendations.

These two restrictions address distinct requirements, and neither implies the other. Equalizing outputs does not establish which recommendations are appropriate: a policy could deliver the same unsuitable list under both cues. Conversely, giving a suitable option the same score under both cues does not ensure that it is generated in both cases or selected under the same rule. DACA therefore brings together two lines of work that address these requirements separately: contrast-targeted learning, which directs a learning objective at a specified contrast rather than at predictive accuracy [12], and representation-based reward debiasing, which trains rewards on representations stripped of spurious information [13]. DACA unites them under a single decision-level objective defined by the same matched cases.

DACA pools the candidates generated under all matched cue variants so that every variant draws on the same options. It trains a cue-invariant reward model against external evidence of quality and aligns final selection. The reward model is frozen during policy optimization, and the resulting policy is judged by a utility measure that is independent of that learned reward. This safeguard matters because consistency alone is easy to obtain: a policy that returns the same list to every student is perfectly consistent across cues yet ignores legitimate differences between students. An update is therefore accepted only if it also meets prespecified utility and responsiveness requirements. Training, policy selection, and held-out testing use disjoint sets of base profiles, limiting adaptation to the evaluation sample [14]. The procedure retains the original policy when no candidate update qualifies.

Our analysis formalizes why the three stages must be governed together. Paired replicates yield unbiased estimates of the cue effect. Stability of the reward model's representation bounds the stability of its scores. Our main result bounds the difference between the recommendation distributions under two cues by the sum of the discrepancies introduced at generation, scoring, and selection. It follows that a common candidate pool, cue-invariant scores, and a cue-independent selector are jointly sufficient for identical distributions. Simple counterexamples show that none

of the three can be omitted from a general guarantee. A final condition specifies when the utility requirement excludes policies that ignore legitimate profile differences.

We evaluate DACA on college-major recommendation, building on Zheng's study of demographic bias in this task [3]. Five open-weight LLMs each return a ranked list of ten majors, drawn from a common catalog of 100, for 20 sets of academic profiles; matched variants change gender, race, or race within a stated gender. Under the original models, changing only the demographic description replaces about 2.5 recommended majors. The complete implementation, which we call *robust* DACA, produces the same majors in the same order for matched variants. Because it satisfies the sufficient conditions of our analysis, this consistency holds by construction; the empirical question is whether it can coexist with better recommendations. The results indicate that it can. Average rubric-based academic fit rises from 63.5 to 86.4, and an index of responsiveness to academic differences remains at its baseline of 100. Omitting the cue from the prompt also yields identical lists, but leaves fit at 63.5 and lowers responsiveness to 91.6. The gain combines consistency with stronger academic fit and preserved responsiveness.

The ablations show how these outcomes arise as a result of individual components. Conventional reward reranking, which learns academic fit without restricting the cue, supplies most of the fit gain and reduces cue-induced movement in the recommended lists by 87.2%. Adding a matched-cue penalty during reward training, which we call aligned reward, closes the remaining gap in STEM exposure. Named majors and their ranks still change across cues, so parity in an aggregate summary does not imply identical recommendations. These residual differences disappear only when a common candidate pool and selection based on a single cue-marginalized score vector per profile are added, two components that the experiment evaluates jointly.

A separate held-out evaluation then tests the policy selected in advance for each model against every prespecified requirement. The responsiveness requirement affects policy selection. For Qwen, validation showed that a cue-invariant update with high fit barely differentiated among academic profiles, so the original policy was retained. Four of the five selected policies meet the cue-stability requirements and one meets every requirement. The remaining policies either fall short on fit or lack the precision needed to certify a limit on how far a policy may depart from the original recommendations. These analyses distinguish the method's average performance from the evidence needed to accept a particular intervention.

Contributions. The paper makes three contributions.

1. **A decision-level method that aligns what is valued with what is delivered.** DACA coordinates a common candidate pool, a cue-invariant reward model, and final-output alignment around the same matched decision case. It couples this alignment target with independently assessed utility. It thereby treats what should be recommended and how consistently it should be delivered as parts of one constrained recommendation problem, rather than treating fairness and usefulness as separate objectives.

2. **A compositional theory of cue effects in recommendation pipelines.** We provide

estimators that separate persistent cue effects from generation noise, connect representation stability to score stability, and bound differences in final recommendation distributions through generation, scoring, and selection. The analysis provides sufficient conditions for exact alignment, shows that omitting control of any one stage precludes a general guarantee, and states when a utility floor excludes policies that ignore legitimate profile differences.

3. **Empirical evidence linking effectiveness, ablation, and acceptance.** Across five models, DACA delivers identical recommendations to matched profiles while improving rubric-based academic fit and preserving aggregate responsiveness. The ablations show that stable aggregate exposure is not evidence of stable recommendations. Prespecified held-out acceptance testing makes the utility and responsiveness requirements consequential, distinguishing an average improvement from approval of a specific model–policy intervention.

2 Related Work

Two streams of literature are related. Fairness criteria and audits specify and measure prohibited cue effects, while mitigation methods constrain a cue's influence on a representation, an output, a reward, or a ranking. For generated recommendations, the gap between them is that the audited cue can act on candidate generation, scoring, and selection at once. Constraining any one of these leaves the delivered recommendation exposed through the others. We review each line of work by what it controls and what it leaves open; Table 1 summarizes the comparison.

Table 1: What each line of related work constrains, and what it leaves open when an LLM generates the recommendation.

Line of work	What it constrains	What remains open
Fairness criteria and counterfactual fairness [15, 16, 17, 18]	Which differences a decision may contain	How to enforce the criterion when the model also generates the options
LLM audits and benchmarks [1, 2, 8, 19]	Measurement of cue sensitivity	How to remove it while improving the recommendation
Fair representations, logit pairing, fair prompting [20, 21, 22, 23]	Cue dependence of one model's representation or output	Which candidates are generated; whether the delivered list is suitable
Reward debiasing and contrast targeting [13, 12]	The evidence used to score, or the contrast to be estimated	Whether consistently valued options are generated and selected
Fair ranking [10]	Exposure over a given candidate set	Options never generated; which options fill equal exposure
Compositional fairness [24, 25, 26]	Whether component guarantees compose	A method and bound for within-profile consistency in a generative pipeline
DACA (this paper)	Candidate pool, reward evidence, and final selection under one matched contrast; utility assessed independently of the reward	Proxies outside the audited contrasts; validity of the external utility criterion

2.1 Defining and Auditing Decision Bias

An observed disparity does not by itself define decision bias; a fairness criterion specifies which differences a decision may contain [27]. Group criteria constrain decision or error rates across groups. For example, demographic parity constrains decision rates across groups; equal opportunity and equalized odds constrain outcome-conditional error rates [15]. In contrast, individual fairness requires that similar individuals receive similar outcomes under a task-specific similarity metric [16]. DACA's target can be read as individual fairness restricted to matched pairs, with the decision standard fixing the metric. Two prompts that share the relevant profile and differ only in the audited cue are at distance zero, so their recommendation distributions should coincide. The matched construction thereby removes the need to learn or elicit a similarity metric.

Counterfactual fairness provides the closest conceptual foundation. It requires a prediction to be invariant to an intervention on a protected attribute within a structural causal model [17], and path-specific criteria separate permissible from impermissible causal pathways [18]. DACA adopts the principle that an attribute's permitted role is specific to the decision, but intervenes on the model's input rather than on the person. The estimand is the model's response to a change in the cue with the relevant profile held fixed, not the causal effect of a person's identity on their qualifications. The target is therefore realizable and testable without a structural model of how identity shapes a profile. Its guarantee, however, covers the audited contrasts rather than every proxy for identity.

Audits of LLMs follow the logic of correspondence experiments, which vary identity signals while holding qualifications fixed [11]. They document cue effects in clinical, employment, educational, and financial advice [1, 2, 3, 4], including effects carried by dialect, names, or institutional prestige rather than explicit labels [8, 2, 9]. Surveys trace such effects to data, representations, training, and inference [28]. For LLM recommenders, the FaiRLLM benchmark evaluates user-side fairness by comparing the similarity of lists generated under neutral and sensitive instructions [19]. Two lessons carry over to mitigation: an audit must cover the forms in which a cue is expressed, and it must compare option identities and ranks rather than a single group summary. Audits, however, measure cue sensitivity; they neither remove it nor specify what an improved recommendation should contain.

2.2 Constraining Representations, Outputs, and Rewards

Most mitigation methods constrain a single object. Fair representation learning limits dependence on a protected attribute while retaining useful information [20], adversarial variants bound the unfairness of downstream predictors [21], and reductions optimize accuracy subject to explicit fairness constraints [29]. For language models, counterfactual logit pairing penalizes changes in a classifier's output when identity terms are substituted [22], the closest precedent for DACA's output penalty. For LLM recommenders, UP5 learns prompts that make sensitive attributes hard to recover from the model's representation while keeping the backbone fixed [23]. Evaluations caution that reducing one association measure need not remove recoverable information or improve other fairness

measures [30, 31]. These methods constrain one model's representation or output for a given input. In a generated recommendation, however, the cue also determines which candidates exist to be scored and selected, so equalizing one stage need not equalize the delivered list. Targets that make an attribute unrecoverable can also discard legitimate information correlated with it, whereas DACA requires only invariance to the cue for a fixed profile.

A second line of work concerns the quality signal that guides a policy. Reinforcement learning from human feedback optimizes a policy against a learned reward [32], so the policy inherits whatever the reward treats as evidence of quality. Ng et al. identify non-spurious latent factors and train the reward on them [13]. Vafa, Athey, and Blei show that representations optimized for prediction can omit information needed to estimate an adjusted disparity, and fine-tune toward that contrast directly [12]. DACA borrows from both but changes the object. Its reward restriction is decision-specific: the score of a given candidate should not change when only the cue changes. Because matched pairs supervise this invariance directly, no identification of latent factors is required. Its contrast is a constraint on model behavior, to be reduced below a tolerance, rather than a quantity to be estimated accurately. The two restrictions are complementary. A cue-invariant reward can value an option consistently without ensuring that it is generated or selected. A contrast penalty can make two lists identical without establishing that either is suitable.

2.3 Composition Across Recommendation Stages

Fair-ranking methods allocate exposure fairly over a given candidate set, typically toward the items being ranked [10]. In a generated recommendation, the candidate set is itself an output that the cue can change. If a suitable option is generated under one cue but not the other, no reranker restricted to the second set can recover it. We call this a *counterfactual support failure*.

Fairness also need not survive composition: individually fair components can form an unfair system [24]. In pipelines where early stages determine the cohort seen by later ones, unfairness can be arbitrarily large even with two stages, naive auditing can miss it, and fairness requires some dependence between the designs of the stages [25]. For multi-component recommenders, Wang et al. provide conditions under which exposure and pairwise-accuracy fairness compose, and find that end-to-end fairness is largely achievable by improving individual components [26]. Our target is different: within-profile consistency of the delivered list rather than group-level exposure. Consistent with Wang et al., restricting the reward in our experiment closes the aggregate STEM-exposure gap; the named majors and their ranks nonetheless still change across cues (Section 5.4).

To our knowledge, no existing method governs candidate availability, scoring evidence, and final selection together against a common counterfactual target, with usefulness assessed by a criterion independent of the learned reward. DACA supplies the cross-stage dependence that the pipeline results call for: the same matched cases define the common candidate pool, restrict the reward model, and align final selection, and an update is accepted only if independently measured utility and responsiveness meet prespecified requirements. The theory bounds differences in delivered recommendations by the three stage discrepancies. It gives sufficient conditions for exact alignment,

none of which can be dropped from a general guarantee (Theorem 4.3 and Corollary 4.4).

3 Decision Setting and Problem Formulation

We formulate debiasing as a constrained recommendation problem. The delivered recommendation should respond to legitimate differences in qualifications, preferences, and constraints, but remain invariant to cues whose influence the decision standard prohibits. The formulation is generic. It requires only that the permitted information, the audited contrasts, a common output encoding, and an external standard of quality be specified in advance. The four subsections define, in turn, the decision case, the cue effect to be removed, the restriction on the reward model, and the constrained problem that combines them. A running example from the college-major task of Section 5 illustrates each element, and Table 2 collects the notation.

3.1 Decision Cases, Relevant Information, and Audited Cues

Running example: college-major recommendation

Decision case. A student asks an LLM for ten suitable college majors, ranked from most to least suitable and drawn from a common catalog of $K = 100$ majors.

Relevant information x_R. Strong mathematics preparation, strong writing preparation, and strong science performance; interests in environmental issues and working with data; a career goal in policy analysis; and a preference for collaborative work.

Audited cue X_A. The student's gender or race description. Two matched prompts share every statement of x_R and all instructions, and differ only in the cue, for example "The student is a woman" versus "The student is a man." Other prespecified contrasts express the cue through pronouns or paraphrase, or change race while holding a stated gender fixed.

Decision standard. Academic fit: x_R may change which majors are recommended and how they are ranked; X_A may not.

A *decision case* is a request for a recommendation, described by two kinds of information. Let X_R denote the *decision-relevant information* that the model is permitted to use, such as qualifications, preferences, and constraints stated by the user. Let X_A denote the *audited cue*: an attribute, an identity-related expression, or a prespecified combination of such expressions whose influence on the recommendation is under test. In the running example, the student's preparation, interests, career goal, and work style form X_R, and the gender or race description forms X_A.

The *decision standard* fixes this partition in advance. It is not inferred from correlations in the data. The standard may derive from law, institutional policy, professional norms, or the purpose of the decision. Title VII, the Age Discrimination in Employment Act, and Title IX, for instance, identify attributes that may not serve as a basis for particular employment and educational decisions. Under the academic-fit standard of the running example, changing the student's interests may legitimately change the recommended majors, whereas changing only the demographic description should not.

Some attributes have both a permitted and a prohibited role. The formulation handles them by placing the authorized information in X_R as an explicit statement, while the identity expression itself remains the audited cue. In financial advice, for example, a user's stated requirement that investments conform to their religious principles is a constraint that may shape the recommendation, whereas a mention of the user's religion, without such a requirement, should not [4]. The target is *within-profile consistency*: the same decision case should receive the same recommendation whatever its audited cue. Cases that differ in X_R may receive different recommendations.

3.2 The Counterfactual Recommendation Effect

The effect to be removed is defined through a *matched comparison*: two versions of the same decision case, (x_R, a) and (x_R, a'), whose prompts share the relevant information, instructions, and output format and differ only in the audited cue. Holding x_R fixed separates the cue effect from legitimate differences between profiles. Testing a specified change in the cue avoids assuming that deleting an attribute removes all associated information. The estimand is thus the model's response to a controlled change in its input, not the causal effect of a person's identity (Section 2.1).

Let π_θ denote the complete *recommendation policy*, comprising the LLM and its rules for generating, scoring, and selecting candidates, with θ indexing the policy. Its delivered output is

$$\boldsymbol{Y}_\theta = \pi_\theta(X_R, X_A, \varepsilon), \qquad \boldsymbol{Y}_\theta \in \mathcal{Y} \subseteq \mathbb{R}^K, \tag{1}$$

where ε represents the policy's generation and selection randomness. The output is expressed in a prespecified, bounded numerical encoding shared by both versions of a case, so that matched recommendations can be compared coordinate by coordinate. For ranked recommendations from a common catalog of K options, a natural choice, used in Section 5, is a *rank-exposure encoding*: coordinate k records the exposure of option k, with a larger weight for a higher rank and zero if the option is not recommended. Omitting an option sets its coordinate to zero, and moving it down the list reduces its weight, so the encoding records both which options are recommended and where they appear.

Because repeated generations can differ for an unchanged prompt, the effect is defined on the policy's expected recommendation, $\bar{\boldsymbol{\pi}}_\theta(x_R, a) := \mathbb{E}_\varepsilon[\pi_\theta(x_R, a, \varepsilon)]$. Let $\mathcal{C}_A$ be the prespecified set of audited contrasts (a, a'), and let P_{audit} be the target distribution over relevant profiles and contrasts. The *matched cue effect* and the *audit risk* are

$$\begin{aligned} \boldsymbol{\tau}_\theta(x_R; a, a') &= \bar{\boldsymbol{\pi}}_\theta(x_R, a) - \bar{\boldsymbol{\pi}}_\theta(x_R, a'), \\ \mathcal{F}(\theta) &= \mathbb{E}_{P_{\text{audit}}}\left[\|\boldsymbol{\tau}_\theta(X_R; a, a')\|_2^2\right]. \end{aligned} \tag{2}$$

Each coordinate of $\boldsymbol{\tau}_\theta$ retains the direction of the change. Under the rank-exposure encoding, a positive coordinate means that the option receives more expected exposure under a than under a' for the same profile. The audit risk squares these within-profile effects before averaging, so that effects of opposite sign in different profiles cannot cancel. Averaging over ε separates a persistent

cue effect from variation between individual responses.

The vector form matters in practice. Suppose the policy always ranks Data Science second for the running-example student under cue a, and Statistics second under cue a', with the other nine majors identical. Both are STEM majors in the same position, so the share of STEM majors and any rank-weighted STEM-exposure summary are unchanged. Yet $\boldsymbol{\tau}_\theta$ has two nonzero coordinates of opposite sign, and the profile contributes positively to $\mathcal{F}(\theta)$. The encoding records the substitutions that a scalar exposure summary hides. Section 5 shows that such substitutions can persist after the STEM-exposure gap has closed.

The target $\mathcal{F}(\theta) = 0$ requires the expected encoded recommendations under a and a' to coincide for almost every audited profile and contrast, while leaving recommendations free to differ across profiles. The contrasts in $\mathcal{C}_A$ are fixed in advance and may be multivalued, intersectional, or alternative textual expressions of the same cue, such as explicit labels, pronouns, and paraphrases. Equality of expected outputs is the basic target. Because two stochastic policies can share a mean yet differ in distribution, Section 4.2 also bounds the distance between the two output distributions. Requiring that distance to be zero gives the individual-fairness reading of Section 2.1.

3.3 Restricted Reward Modeling

A zero cue effect does not say which recommendations should be delivered. Mitigation therefore also requires a *reward model* that assesses how well a candidate fits the relevant profile without treating the audited cue as evidence of merit. The two requirements are distinct. Suppose the reward model undervalues Data Science for the running-example student under one cue, despite unchanged quantitative preparation and interests. Excluding Data Science under both cues would satisfy $\mathcal{F}(\theta) = 0$ without establishing that the exclusion is justified. Conversely, giving Data Science the same score under both cues does not ensure that it is generated in both cases or selected by the same rule.

For a candidate recommendation y, the reward model consists of an encoder f_ϕ and a scorer r_ψ, with parameters ϕ and ψ:

$$(\boldsymbol{Z}_R, \boldsymbol{Z}_A) = f_\phi(X_R, X_A, y), \qquad \widetilde{R}_{\phi,\psi}(X_R, X_A, y) = r_\psi(\boldsymbol{Z}_R). \tag{3}$$

The *task-relevant representation* $\boldsymbol{Z}_R$ is intended to carry evidence about the fit between y and X_R, while the *cue representation* $\boldsymbol{Z}_A$ absorbs variation associated with the audited cue. Only $\boldsymbol{Z}_R$ enters the scorer, so the *restricted reward* $\widetilde{R}_{\phi,\psi}$ can depend on the cue only through cue information that remains in $\boldsymbol{Z}_R$. Section 4 specifies how the two representations are estimated.

The restriction is measured on matched candidates. Let P_{pair} be a distribution over tuples (X_R, a, a', y) in which the same candidate y is drawn from a pool shared by the two versions of the case and evaluated under both cue values. Writing $\boldsymbol{Z}_R(a)$ for the task-relevant representation at cue a, with X_R and y held fixed, the *separation loss* is

$$\mathcal{L}_{\text{sep}}(\phi) = \mathbb{E}_{P_{\text{pair}}}\left[\|\boldsymbol{Z}_R(a) - \boldsymbol{Z}_R(a')\|_2^2\right]. \tag{4}$$

It penalizes any change in the evidence used to assess an identical recommendation when only the cue changes. Holding both the profile and the candidate fixed is essential, since comparing different candidates would mix cue sensitivity with genuine differences in fit. The shared pool makes every candidate assessable under both cues, preventing the counterfactual support failure of Section 2.3; Section 4 describes its construction.

The restriction is conditional on the relevant profile. It does not require the task-relevant representation to be independent of demographics across the population. Academic preparation can remain informative about fit even when its distribution differs across demographic groups. The aim is to remove the effect of changing the cue while preserving the information in preparation, interests, and goals. The reward model must therefore also predict fit accurately; invariance alone is not its objective.

3.4 The Constrained Recommendation Problem

The two requirements are combined with a safeguard on usefulness. Let $\mathcal{L}_{\text{quality}}(\theta, \phi, \psi)$ measure the shortfall of the recommendations relative to external quality evidence, such as blinded expert judgments, validated outcomes, or a prespecified rubric. It depends on (ϕ, ψ) through the reward model's assessment of candidates and on θ through the recommendations the policy delivers. Let $U(\pi_\theta)$ measure the utility of the delivered recommendations, assessed independently of the learned reward. Let $u_\star$ be the required utility level, fixed in advance either as an absolute floor or relative to the original policy π_0, as $u_\star = U(\pi_0) - \delta$. The population target is

$$\begin{aligned} \min_{\theta,\phi,\psi} \quad & \mathcal{L}_{\text{quality}}(\theta, \phi, \psi) \\ \text{subject to} \quad & \mathcal{L}_{\text{sep}}(\phi) \le \xi_{\text{rep}}, \\ & \mathcal{F}(\theta) \le \xi_{\text{cf}}, \\ & U(\pi_\theta) \ge u_\star. \end{aligned} \tag{5}$$

The nonnegative tolerances ξ_{rep} and ξ_{cf} bound the representation discrepancy and the final-output cue effect; setting either to zero gives its exact invariance target. Under a relative requirement, the nonnegative tolerance δ is the largest acceptable loss of utility relative to π_0.

The objective favors high-quality recommendations. The first two constraints require that their assessment and delivery comply with the decision standard, and the last makes independently measured usefulness a condition of acceptance. The last constraint is not redundant. A policy that returns one fixed list to every student satisfies $\mathcal{F}(\theta) = 0$ exactly but ignores differences in preparation and interests. The utility measure must therefore reward fit to legitimate profile information, and $u_\star$ must be high enough to exclude such profile-insensitive policies; Section 4.2 provides a sufficient condition. Applications may add further prespecified requirements of the same form, such as a minimum responsiveness to differences in X_R or a limit on how far the policy may depart from the original recommendations; Section 5 imposes both, together with a candidate-coverage requirement.

The learned reward guides optimization, whereas external utility assesses what optimization

produces. Evaluating utility independently prevents a policy from qualifying merely because it scores well under its own reward model. In the running example, utility is the rubric-based academic fit of the delivered ranking, computed on profiles not used to fit the reward model. The objective and constraints require recommendations to fit permissible evidence and remain stable under the audited change of cue.

Table 2: Notation used in the formulation and method.

Symbol	Meaning
X_R, X_A	Decision-relevant information and the audited cue.
(a, a'), $\mathcal{C}_A$, P_{audit}	A matched cue contrast, the prespecified set of contrasts, and the target distribution over profiles and contrasts.
π_θ, $\boldsymbol{Y}_\theta$, $\bar{\boldsymbol{\pi}}_\theta$, ε	Complete recommendation policy, its encoded output in $\mathbb{R}^K$, its expected output, and its generation and selection randomness.
$\boldsymbol{\tau}_\theta$, $\mathcal{F}$	Matched cue effect and audit risk, its mean squared magnitude.
f_ϕ, r_ψ, $\boldsymbol{Z}_R$, $\boldsymbol{Z}_A$	Encoder, scorer, task-relevant representation, and cue representation.
$\widetilde{R}_{\phi,\psi}$, P_{pair}	Restricted reward and the matched-candidate distribution.
$\mathcal{L}_{\text{sep}}$, $\mathcal{L}_{\text{quality}}$	Separation loss and shortfall relative to external quality evidence.
U, $u_\star$, π_0, δ	External utility, its required level, the original policy, and the tolerance of a relative requirement.
ξ_{rep}, ξ_{cf}	Tolerances for representation discrepancy and final-output cue effect.
B	Paired generation replicates per profile and contrast (Section 4).

4 Decision-Aware Counterfactual Alignment

DACA solves the problem set up in Section 3 in stages. Its guiding observation is that the audited cue can reach the delivered recommendation by three routes: it can change which options are generated, how they are scored, and which are selected and in what order. In the running example, a suitable major such as Data Science may be missing from the candidates generated under one demographic description, receive a lower score under that description, or be ranked lower by the final selection rule. Correcting its score closes only the second route. DACA therefore coordinates a common candidate pool, a restricted reward model, and final-output alignment around the same matched cases. It adopts the result only if independently measured utility remains adequate. Section 4.1 describes the components and why each is needed, Section 4.2 states the results that justify them, and Section 4.3 assembles the complete procedure.

4.1 Five Components and Their Roles

Figure 1 organizes the method into five components. C1 constructs matched decision cases. C2, C3, and C4 each close one route: C2 makes the same options available under every cue, C3 trains a reward model that scores candidates on task-relevant evidence, and C4 aligns the final recommendation. C5 decides whether the resulting policy is useful enough to adopt. The components correspond to Equation (5). C3 fits the reward model with the separation constraint as a penalty. C4 fits the policy

against the frozen reward, with the cue-effect constraint as a penalty and the utility requirement as a constraint; C5 checks every requirement at its prespecified tolerance. DACA can operate around a fixed generator through candidate pooling and external reranking, or train accessible parts of the recommendation system directly.

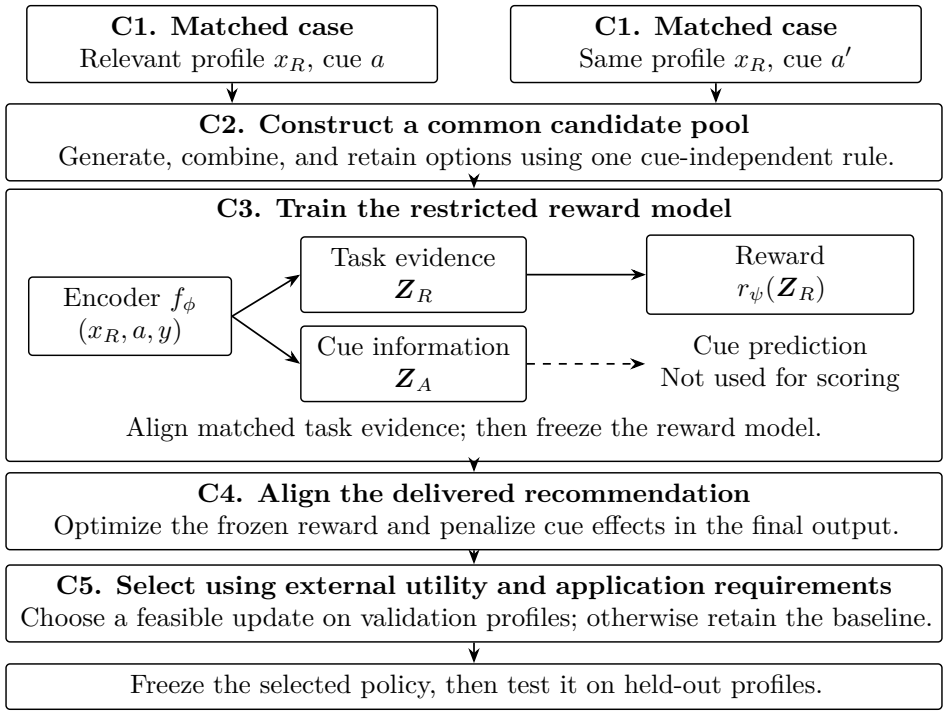


Fig. 1 The DACA pipeline. Matched cases (C1) share the relevant profile and a common candidate pool (C2). Only the task-relevant representation enters the reward (C3). Final-output alignment (C4) and external utility assessment (C5) govern the resulting policy. C2, C3, and C4 target the three stage discrepancies of Theorem 4.3 (Table 3); Table 4 maps the components to the empirical ablations

C1: Construct matched cases. For each relevant profile x_R and contrast $(a, a') \in \mathcal{C}_A$, DACA constructs two prompts that differ only in the audited cue and collects B generations under each. Matching removes legitimate differences between profiles from the comparison, and replication separates a persistent cue effect from ordinary sampling variation (Theorem 4.1). All variants of a base profile are kept in the same training, validation, or held-out partition. Later evaluation therefore concerns new profiles rather than new cue variants of profiles seen during development.

C2: Provide a common candidate pool. A reward model can assess only the candidates made available to it. No reranker can recover an option that was never generated. This is the counterfactual support failure described in Section 2.3. DACA therefore pools candidates across the matched variants before any scoring:

$$\overline{\mathcal{C}}(x_R) = \bigcup_{v \in \mathcal{A}_A} \bigcup_{b=1}^{B} G(x_R, v, \varepsilon_{vb}), \tag{6}$$

where G is the candidate generator, $\mathcal{A}_A$ contains the cue values appearing in $\mathcal{C}_A$, and ε_{vb} is the randomness of the bth generation under cue v. Duplicates are removed and, if a capacity limit is needed, a cue-independent retention rule is applied once; every variant of the profile then uses the same realized pool. Pooling makes availability comparable without forcing any option into the recommendation. In the running example, a major generated under only one description becomes available under both, but whether it is recommended still depends on its academic fit. When model weights are accessible, the generator can instead be trained toward common support.

C3: Train a restricted reward model. Let $q^\star(x_R, y)$ denote external evidence of the quality of candidate y for profile x_R, obtained from cue-blinded judgments, validated outcomes, or a prespecified rubric. The encoder and scorer of Equation (3) are fitted on matched training tuples by

$$(\widehat{\phi}, \widehat{\psi}) \in \arg\min_{\phi,\psi} \; \mathcal{L}_Q(\phi, \psi) + \lambda_{\text{rep}} \mathcal{L}_{\text{sep}}(\phi) + \lambda_{\text{aux}} \mathcal{L}_{\text{aux}}(\phi), \tag{7}$$

where $\mathcal{L}_Q(\phi, \psi) = \mathbb{E}\, \ell\{\widetilde{R}_{\phi,\psi}(X_R, X_A, y), q^\star(X_R, y)\}$ for a predictive or preference loss ℓ, and the same profile and candidate are held fixed within each contrast. Each term has a distinct job. The quality loss $\mathcal{L}_Q$, the reward-model component of $\mathcal{L}_{\text{quality}}$ in problem (5), keeps $\boldsymbol{Z}_R$ informative about fit. The separation loss $\mathcal{L}_{\text{sep}}$ of Equation (4) removes cue-induced change from that evidence. The auxiliary loss $\mathcal{L}_{\text{aux}}$, implemented for example as a cue-prediction head on $\boldsymbol{Z}_A$, gives cue information a separate representation that the scorer cannot access. This reduces the encoder's incentive to store it in $\boldsymbol{Z}_R$. The nonnegative weights λ_{rep} and λ_{aux} set their emphasis. Removing the separation penalty yields an unrestricted reward model, which corresponds to the conventional-reward comparison of Section 5.

After fitting, the reward model is frozen. Freezing keeps the criterion of quality fixed while the policy is optimized against it. The policy therefore cannot raise its score by changing what the score measures. Utility in C5 is then assessed against external evidence rather than this learned reward.

C4: Align the final recommendation. With the reward model frozen, DACA learns the recommendation policy by reranking, fine-tuning, or both. Let $\widehat{Q}(\theta) = \widehat{\mathbb{E}}\, \widetilde{R}_{\widehat{\phi},\widehat{\psi}}(X_R, X_A, \boldsymbol{Y}_\theta)$ be the mean frozen reward of the policy's outputs on training profiles. The reward of a ranked list aggregates its candidates' rewards by a fixed task-specific rule. Let $\widehat{\mathcal{F}}(\theta)$ be the empirical audit risk, obtained by averaging a per-case estimate of $\|\boldsymbol{\tau}_\theta\|_2^2$ from Theorem 4.1 over training profiles and contrasts. The policy solves

$$\widehat{\theta} \in \arg\min_{\theta} \; -\widehat{Q}(\theta) + \lambda_{\text{cf}} \widehat{\mathcal{F}}(\theta) \quad \text{subject to} \quad \widehat{U}(\pi_\theta) \geq \widehat{u}_\star, \tag{8}$$

where $\widehat{u}_\star$ is the empirical counterpart of the required utility level $u_\star$: a fixed absolute floor, or $\widehat{U}(\pi_0) - \delta$ for a relative requirement. The first term favors suitable recommendations; the second penalizes matched differences in the delivered output, including changes of option and, when encoded, of rank. The weight $\lambda_{\text{cf}} \geq 0$ sets the emphasis on alignment.

C4 is needed even when C3 succeeds, because scores that are equal on average do not guarantee equal lists. Lemma 4.2 controls score differences only on average, and a small difference near the selection boundary can still change which option enters the list or where it appears. C4 therefore acts on the delivered recommendation itself. It can be implemented through the penalty in Equation (8) or by construction. Scoring each pooled candidate once per profile, by marginalizing its frozen reward over the matched cue values, and selecting with a cue-independent rule gives every variant identical scores and an identical selector. The experiments in Section 5 use this construction, called *robust* DACA. C3 remains necessary under marginalization, because it determines what the common score reflects, whereas marginalization only makes the score common. A reward that undervalued Data Science under one cue would, once marginalized, undervalue it for every student.

C5: Select an acceptable recommendation rule. Cue invariance is necessary but not sufficient for adoption. On validation profiles, DACA checks each candidate policy against the prespecified requirements: representation separation, final-output alignment, utility of at least $u_\star$ under the external criterion rather than the learned reward, and any additional application requirements. It then selects the feasible policy with the highest external utility. Hyperparameters such as λ_{rep} and λ_{cf} guide learning, whereas the tolerances determine acceptability. The chosen policy and evaluation protocol are frozen before held-out testing, limiting adaptation to the test sample [14]. If no candidate qualifies, the original policy is retained without certification. With development, selection, and evaluation kept separate, a no-update outcome is a legitimate result of the procedure.

4.2 Theoretical Properties and Their Interpretation

The results answer four questions, one for each part of the design: how to estimate a persistent cue effect from stochastic generations (C1), when a stable representation implies a stable score (C3), how the three stages jointly determine the delivered recommendation (C2–C4), and when the utility requirement excludes uninformative solutions (C5). All results concern the controlled prompt contrast defined in Section 3, and Appendix A proves them in the same order.

Estimating a cue effect rather than generation noise. A single pair of responses mixes the cue effect with sampling noise, because a stochastic LLM can produce different lists for an unchanged prompt. The target is the systematic change in the expected recommendation. For a fixed profile x_R and contrast (a, a'), draw B replicate pairs and define

$$\widehat{\boldsymbol{\tau}}^{(b)} = \pi_\theta(x_R, a, \varepsilon_b) - \pi_\theta(x_R, a', \varepsilon'_b), \qquad \widehat{\boldsymbol{\tau}}_B = \frac{1}{B}\sum_{b=1}^{B} \widehat{\boldsymbol{\tau}}^{(b)}, \qquad \widehat{\sigma}_B^2 = \frac{1}{B-1}\sum_{b=1}^{B} \|\widehat{\boldsymbol{\tau}}^{(b)} - \widehat{\boldsymbol{\tau}}_B\|_2^2. \quad (9)$$

Each $\widehat{\boldsymbol{\tau}}^{(b)}$ is a single-replicate estimate of $\boldsymbol{\tau}_\theta(x_R; a, a')$, $\widehat{\boldsymbol{\tau}}_B$ averages them, and $\widehat{\sigma}_B^2$ is their total sample variance. The two responses within a pair may be drawn independently or coupled, for example through a shared seed, provided each follows its correct policy distribution.

Theorem 4.1 (Matched-prompt estimation)**.** *Suppose the replicate pairs are independent and identically distributed and each response follows its policy distribution. Write* $\boldsymbol{\tau} = \boldsymbol{\tau}_\theta(x_R; a, a')$.

(i) Mean effect. $\mathbb{E}(\widehat{\boldsymbol{\tau}}_B) = \boldsymbol{\tau}$, *and* $\widehat{\boldsymbol{\tau}}_B \to \boldsymbol{\tau}$ *almost surely as* $B \to \infty$.

(ii) Squared effect. *For* $B \geq 2$, *the corrected estimator*

$$\widehat{f}_B = \|\widehat{\boldsymbol{\tau}}_B\|_2^2 - \widehat{\sigma}_B^2/B \tag{10}$$

satisfies $\mathbb{E}(\widehat{f}_B) = \|\boldsymbol{\tau}\|_2^2$.

(iii) Precision. *If every coordinate of* $\widehat{\boldsymbol{\tau}}^{(b)}$ *lies in a fixed interval of width at most* w, *then for* $0 < \alpha < 1$, *with probability at least* $1 - \alpha$,

$$\|\widehat{\boldsymbol{\tau}}_B - \boldsymbol{\tau}\|_\infty \leq w\sqrt{\frac{\log(2K/\alpha)}{2B}}. \tag{11}$$

Part (i) says that the matched average targets the intended effect. Part (ii) addresses overestimation from squaring a noisy average, because

$$\mathbb{E}\|\widehat{\boldsymbol{\tau}}_B\|_2^2 = \|\boldsymbol{\tau}\|_2^2 + \sigma_\theta^2/B, \qquad \sigma_\theta^2 = \mathbb{E}\|\widehat{\boldsymbol{\tau}}^{(1)} - \boldsymbol{\tau}\|_2^2,$$

and the excess σ_θ^2/B is simulation variance, not a cue effect. The corrected estimator subtracts an unbiased estimate of that excess; equivalently, it averages the inner products $\widehat{\boldsymbol{\tau}}^{(b)\top}\widehat{\boldsymbol{\tau}}^{(b')}$ between distinct replicates, each of which has expectation exactly $\|\boldsymbol{\tau}\|_2^2$. Averaging $\widehat{f}_B$ over profiles and contrasts drawn from P_{audit} therefore estimates the audit risk $\mathcal{F}(\theta)$ without variance inflation. The uncorrected plug-in $\|\widehat{\boldsymbol{\tau}}_B\|_2^2$ remains usable as a nonnegative training penalty in C4, but because its excess depends on θ, penalizing it also rewards lower generation variance. Part (iii) bounds the worst coordinate error simultaneously over all K options. The bound shrinks at rate $B^{-1/2}$, so quadrupling the number of replicate pairs halves it. The bounded encoding of Section 3.2 supplies the width w. The theorem thus links the repetition budget in C1 to the precision of the reported cue effect.

From a stable representation to a stable score. C3 penalizes changes in a representation, but selection uses scores. The next result relates the two under a Lipschitz condition on the scorer: r_ψ is L_R-Lipschitz if $|r_\psi(\boldsymbol{z}) - r_\psi(\boldsymbol{z}')| \leq L_R\|\boldsymbol{z} - \boldsymbol{z}'\|_2$. The constant L_R caps how much the scorer can amplify a change in its input.

Lemma 4.2 (Restricted-score stability)**.** *If* r_ψ *is* L_R*-Lipschitz, then*

$$\mathbb{E}_{P_{\text{pair}}}\left[\left|\widetilde{R}_{\phi,\psi}(X_R, a, y) - \widetilde{R}_{\phi,\psi}(X_R, a', y)\right|^2\right] \leq L_R^2\,\mathcal{L}_{\text{sep}}(\phi). \tag{12}$$

The lemma relates representation stability to score stability: $\mathcal{L}_{\text{sep}} \leq \xi_{\text{rep}}$ bounds the root-mean-square change in a candidate's score by $L_R\sqrt{\xi_{\text{rep}}}$, and zero separation loss yields equal scores almost

surely under P_{pair}. Two limits matter for the design. The guarantee is an average rather than a uniform bound over candidates, and it concerns scores, not the options that are generated or selected. Carrying score stability through to the recommendation therefore requires C2 and C4, which the next result makes precise.

Bounding the cue effect across the complete pipeline. Fix a profile and a contrast. Under cue a, let C_a be the generated candidate pool, $\boldsymbol{s}_a(C)$ the vector of restricted scores assigned to the candidates in a pool C, and T_a the rule that selects and orders the final recommendation using selection randomness η. To see how each stage contributes, the analysis moves from the pipeline under a to the pipeline under a' one stage at a time (Table 3): first the scores are swapped, then the selector, then the candidate pool. Each swap changes the output by an expected amount Δ, measured in the units of the encoded recommendation. The intermediate outputs are analytical devices, not deployed policies. The deployed pipeline still generates, scores, and selects in that order. Let P_a and $P_{a'}$ be the output distributions under the two cues, and let W_1 denote the Wasserstein distance: the smallest expected distance between draws from the two distributions over all couplings that preserve them.

Table 3: Stage-by-stage path from the pipeline under a to the pipeline under a'. Each row changes one stage of the row above; the profile and the selection randomness η are held fixed. The discrepancy is the expected change in the encoded output at that step, and the last column names the DACA component that targets it.

Output	Definition	Stage changed	Discrepancy	Targeted by
$\boldsymbol{Y}_0$	$T_a\{C_a,\, \boldsymbol{s}_a(C_a),\, \eta\}$	None: pipeline under a	–	–
$\boldsymbol{Y}_1$	$T_a\{C_a,\, \boldsymbol{s}_{a'}(C_a),\, \eta\}$	Scoring	$\Delta_{\text{score}} = \mathbb{E}\|\boldsymbol{Y}_0 - \boldsymbol{Y}_1\|_2$	C3
$\boldsymbol{Y}_2$	$T_{a'}\{C_a,\, \boldsymbol{s}_{a'}(C_a),\, \eta\}$	Selection	$\Delta_{\text{select}} = \mathbb{E}\|\boldsymbol{Y}_1 - \boldsymbol{Y}_2\|_2$	C4
$\boldsymbol{Y}_3$	$T_{a'}\{C_{a'},\, \boldsymbol{s}_{a'}(C_{a'}),\, \eta\}$	Availability: pipeline under a'	$\Delta_{\text{pool}} = \mathbb{E}\|\boldsymbol{Y}_2 - \boldsymbol{Y}_3\|_2$	C2

Theorem 4.3 (End-to-end counterfactual bound)**.** *Suppose the outputs in Table 3 are constructed on a common probability space with $\boldsymbol{Y}_0 \sim P_a$ and $\boldsymbol{Y}_3 \sim P_{a'}$. Then*

$$\|\boldsymbol{\tau}_\theta(x_R; a, a')\|_2 \le W_1(P_a, P_{a'}) \le \Delta_{\text{score}} + \Delta_{\text{select}} + \Delta_{\text{pool}}, \tag{13}$$

and consequently $\mathcal{F}(\theta) \le \mathbb{E}_{P_{\text{audit}}}[(\Delta_{\text{score}} + \Delta_{\text{select}} + \Delta_{\text{pool}})^2]$.

The right inequality is the triangle inequality along the path of swaps: the two endpoints can be no farther apart than the total distance traveled between them. The left inequality says that a small distance between output distributions forces a small mean effect, so the bound also controls $\boldsymbol{\tau}_\theta$. The theorem has two implications for the design. First, the delivered recommendation can differ across cues through any stage, so controlling one stage leaves the other two open: stable scores leave availability and selection, and a common pool leaves how the same options are scored and ranked. Each DACA component targets one term, as the last column of Table 3 shows, which is the formal reason for coordinating C2–C4. Section 5 illustrates this implication: with aligned reward, matched scores are nearly identical, yet the recommended lists still differ until the common pool

and cue-marginalized selection are added together. Second, the middle term compares complete output distributions, so controlling the stage discrepancies also controls differences in stochastic behavior that $\mathcal{F}(\theta)$ alone would miss. The bound is not an additive decomposition of the observed effect. The Δ terms depend on the coupling and the order of swaps, and the bound holds for any prespecified choice.

Why small score differences can still change a ranked list. Score changes matter most near a selection boundary. Suppose selection is deterministic top-k on a common pool and every score changes by at most κ across cues, so that $\|\boldsymbol{s}_a - \boldsymbol{s}_{a'}\|_\infty \leq \kappa$. If the gap between the kth and $(k+1)$st scores exceeds 2κ, no included candidate can fall below an excluded one, and the selected set is unchanged. Preserving the order within the set further requires gaps greater than 2κ between adjacent selected candidates. When these margins hold, $\Delta_{\text{score}} = 0$ for a rank encoding. The condition differs from Lemma 4.2 in two ways: it requires a uniform rather than an average bound on score changes, and it depends on the score margins among the profile's candidates, which the separation loss does not control. This is why final-output alignment remains useful when matched scores are already close.

Conditions for exact alignment. Setting every term of Theorem 4.3 to zero gives a sufficient condition for identical output distributions. We call the resulting property *counterfactual closure*: no cue effect remains in the composition of generation, scoring, and selection.

Corollary 4.4 (Counterfactual closure)**.** *Suppose that, for P_{audit}-almost every profile and contrast, matched variants use the same realized candidate pool, assign every candidate the same score, and apply the same cue-independent selector with shared randomness. Then $P_a = P_{a'}$, $\boldsymbol{\tau}_\theta(x_R; a, a') = 0$, and $\mathcal{F}(\theta) = 0$. None of the three conditions can be omitted from a guarantee that places no further restrictions on the stages.*

The conditions are jointly sufficient, not necessary for every invariant policy. They secure invariance without relying on effects at different stages happening to cancel. Their nonredundancy is visible with two options u and v and a rule that selects one. Different pools, $\{u\}$ versus $\{v\}$, force different choices under identical scores and selectors; different scores, $(1, 0)$ versus $(0, 1)$ on a common pool, reverse the choice; and cue-dependent tie-breaking changes the choice even with a common pool and identical scores. When the conditions hold and selection is deterministic, matched variants receive the same recommendation; with stochastic selection, shared randomness makes paired realizations identical, whereas independent draws from the same distribution can still differ. Robust DACA satisfies all three conditions by construction, which is why its consistency in Section 5 is exact. The guarantee remains conditional on the relevant profile and therefore permits personalization.

Excluding profile-insensitive policies. Closure governs differences across cues but does not assess recommendation quality. For example, a policy that always returns the same list is closed

trivially while ignoring the student's preparation and interests. Let Π_{flat} contain the policies whose output distribution does not respond to X_R, and let $U_{\text{flat}} = \sup_{\pi \in \Pi_{\text{flat}}} U(\pi)$ be the best utility such a policy can achieve. If $u_\star > U_{\text{flat}}$, no profile-insensitive policy satisfies the utility constraint in problem (5); for a relative requirement, the condition reads $U(\pi_0) - \delta > U_{\text{flat}}$. When utility rewards profile-specific fit and its floor exceeds what a generic list can achieve, C5 prevents alignment from being obtained by abandoning personalization. It also suggests how to set the floor: for a utility that averages over recommendations, U_{flat} is the utility of the best fixed list. Section 5 enforces the same idea directly through a minimum profile-response requirement.

4.3 Development, Policy Selection, and Held-Out Testing

Algorithm 1 assembles the components into one procedure that separates training, selection, and testing, for both external reranking and trainable implementations. The application choices, namely the profile generator, audited contrasts, output encoding, utility standard, and tolerances, are fixed before fitting. Section 5 specifies them for the college-major task.

Inputs: original policy π_0; base profiles; contrasts $\mathcal{C}_A$; replicate count B; external quality evidence $q^\star$; utility U and required level $u_\star$; tolerances ξ_{rep} and ξ_{cf}; any additional application requirements.

1. **Prespecify and split.** Fix the task, output encoding, requirements, and contrasts. Split base profiles into training, validation, and held-out sets, keeping all variants of a profile together.
2. **Match and generate [C1].** Construct the cue variants and collect B paired replicates under the intended policy distributions.
3. **Pool [C2].** Combine and deduplicate candidates, apply any capacity rule once, and reuse the resulting pool for every variant of a profile.
4. **Fit and freeze the reward model [C3].** Solve Equation (7) on training profiles, then freeze the encoder and scorer.
5. **Construct candidate policies [C4].** Rerank, fine-tune, or both, using Equation (8) or cue-marginalized selection.
6. **Select [C5].** On validation profiles, choose the highest-utility policy that meets every requirement; if none qualifies, retain π_0 without certification.
7. **Freeze and test.** Fix the chosen policy and evaluation protocol, and evaluate the frozen choice on held-out profiles against the prespecified requirements; do not reselect using test outcomes.

Output: the selected recommendation rule and its held-out assessment, or a no-update outcome.

Algorithm 1: DACA development and evaluation

5 Empirical Validation: Experiment Design and Results

We evaluate DACA on college-major recommendation, a task in which a useful policy must distinguish students with different preparation and interests while giving the same advice when only a demographic description changes. The evaluation answers four questions, each tied to a result of

Section 4.2: (i) whether the complete method removes cue effects while improving fit and preserving responsiveness; (ii) how reward modeling and complete-output alignment contribute; (iii) whether consistency holds within each type of comparison and form of cue; and (iv) whether the rule selected in advance for each model meets every prespecified requirement on new profiles. Across five models, robust DACA gives matched profiles identical lists, raises average academic fit from 63.5 to 86.4, and preserves responsiveness to academic differences. It is the only policy tested that attains the best value on every measure of cue discrepancy and fit.

5.1 Task, Models, and Audited Contrasts

From documented bias to a mitigation task. Zheng [3] examined GPT-3.5-Turbo's college-major recommendations for profiles built from public 2023 California Assessment of Student Performance and Progress data. Prompts described twelfth-grade test-score percentiles and demographic characteristics and requested ten suitable majors, including both STEM and non-STEM fields. With academic information held fixed, the recommended options and their rank-weighted STEM exposure differed across demographic descriptions. At the 50th score percentile, STEM Disparity Scores were 0.98, 0.77, and 0.51 for profiles labeled male, female, and LGBTQ+, with further differences by race and socioeconomic status. The present experiment asks whether such recommendations can be made consistent across demographic descriptions while improving their fit to academic information.

Profiles, catalog, and contrasts. The experiment uses 20 most prevalent families of academic profiles, of which the running example of Section 3 is one instance. The relevant information X_R comprises interests, career goals, work style, mathematics and writing preparation, and science performance; gender and race descriptions form the audited cue X_A under the academic-fit standard. Each response ranks ten distinct majors from a common catalog of 100, so that changes in both membership and rank are observable. The contrast set $\mathcal{C}_A$ contains ten prespecified contrasts of three types: *gender* (woman versus man and nonbinary versus man), each expressed through explicit labels and through pronouns; *race* (Asian versus White and Black versus White), each expressed through explicit labels and through paraphrase; and *race within gender* (Asian versus Black at a stated gender of woman or of man). Each family also has a version without any demographic description, which the cue-omission baseline uses.

Models. We use five locally served open-weight models: Qwen 3.5 4B, GPT-OSS 20B, DeepSeek-R1 8B, Gemma 3 4B, and Llama 3.1 8B. Open weights are needed because DACA requires control over response collection, reward-model training, and reproducible comparison of training and inference stages. A commercial inference API does not provide this control. The study therefore evaluates the method on these models rather than reproducing Zheng's GPT-3.5-Turbo setting, which can be prompted but not retrained. We use *open-weight* to describe this access, without a further claim about each model's licensing terms.

5.2 Evaluation Scheme

Data partition. The 20 profile families were partitioned before modeling into 8 training, 4 validation, and 8 held-out families, with all cue variants of a family kept in the same partition. Training families fit the reward models, validation families select one recommendation rule per model (C5), and held-out families evaluate the frozen choice. Held-out testing therefore concerns new academic profiles, not merely new demographic descriptions of profiles seen during development.

Policies compared. Each policy is a complete recommendation rule for a given model.

- *Original generation* returns the model's own ranked list.
- *Cue omission* returns the list generated from the version of the profile without a demographic description.
- *Conventional reward* reranks each variant's own candidates by a reward model fitted without the separation penalty.
- *Aligned reward* performs the same reranking with the separation penalty of C3.
- *Robust* DACA adds a common candidate pool (C2) and selects from one cue-marginalized score vector per profile (C4 by construction).

The reward models are fitted on training families against rubric-based assessments of candidate majors. Utility is computed from the delivered rankings on held-out families, not from the reward models' scores. This keeps utility independent of the learned reward in the sense of Section 3.4.

Two analyses on the held-out panel. The *common-policy comparison* applies each of the five policies to every model and averages the results equally across models. It identifies what each component contributes (Sections 5.4 and 5.5). *Model-specific acceptance* tests only the rule selected for each model on validation profiles, against the full set of prespecified requirements and using simultaneous confidence bounds (Section 5.6). The first analysis explains how the method works; the second decides whether a particular intervention may be adopted. Table 4 maps each analysis to its question and to the theoretical result it tests.

Measures. Cue discrepancy is measured in three ways. *Movement* ($100M$) is a transport discrepancy between matched ranked lists; it is zero only when they contain the same majors in the same order. The *STEM gap* is the absolute difference in rank-weighted STEM exposure, in percentage points. The *exposure discrepancy* ($100E$) is the Euclidean discrepancy between matched rank-exposure vectors over all 100 majors (Section 3.2). The STEM gap is a scalar summary, whereas movement and exposure discrepancy also register substitutions that leave it unchanged, as in the Data Science and Statistics example. We assess usefulness with three measures. *Academic fit* ($100U$) is the normalized discounted cumulative gain against the prespecified academic-fit rubric [33]. A value of 100 denotes the rubric's ideal ranking rather than a percentage of correct advice. *Lowest-cue*

Table 4: Evaluation design. Every analysis uses the held-out panel and the matched cases of C1; the last column names the result of Section 4.2 that the analysis tests.

Analysis (section)	Policies compared	Question	Result tested
Joint effectiveness (5.4)	Robust DACA versus original generation and cue omission	Is consistency accompanied by better fit and preserved responsiveness?	Corollary 4.4; utility floor
Reward modeling, C3 (5.4)	Original generation, conventional reward, aligned reward	What do learning fit, and restricting the cue in the reward, change?	Lemma 4.2
Complete-output alignment, C2 and C4 (5.4)	Aligned reward versus robust DACA	Are the remaining substitutions and rank changes removed?	Theorem 4.3; Corollary 4.4
Comparison types and cue forms, C1 (5.5)	Original generation versus robust DACA, by contrast	Does consistency hold within each comparison type and wording?	Corollary 4.4, per contrast
Selection and acceptance, C5 (5.6)	Rule selected on validation profiles, tested on held-out profiles	Does the selected rule meet every prespecified requirement?	Utility floor

fit ($100U_{\min}$) is the fit of the least favorably treated cue variant. *Profile response* measures how much recommendations differ across legitimately different academic profiles. The comparison tables report profile response as an index, $100P/P_0$, where P_0 is the absolute response of original generation. Values above 100 therefore mean that a policy differentiates among profiles more than the original does. Acceptance tests use the absolute statistic P. *Source distance* ($100T$) measures how far a policy revises the original recommendations; it describes the extent of the intervention, not its quality. Appendix B.2 collects all scales, including candidate coverage.

What the analysis predicts. Section 4.2 yields three predictions. First, robust DACA satisfies the conditions of Corollary 4.4 by construction, through its common pool and cue-marginalized selection. Cue omission satisfies them trivially because deleting the cue makes matched prompts identical. Both should therefore show zero cue discrepancy, and the informative comparison between them concerns fit and responsiveness. Second, by Theorem 4.3, restricting the reward alone need not equalize the delivered lists. Residual differences should persist, even when matched scores and aggregate exposure are equalized, if candidate pools or selection still differ across cues or small score differences remain near the selection boundary. Third, by the utility floor, consistency does not guarantee useful advice, so a consistent policy may still fail the fit or responsiveness requirements.

5.3 Cue Sensitivity Before Mitigation

All five models change their recommendations when only the demographic description changes, but to very different degrees. A separate motivating audit of 4,250 recommendations, covering gender identity, sexual orientation, socioeconomic status, and race or ethnicity (Appendix B.1), characterizes each model before mitigation (see Section 5.6). DeepSeek-R1 8B is the most cue-

sensitive. Its matched lists share only 4.66 of ten majors, so more than half of each list changes. Both its movement (36.2) and its STEM shift (6.1 percentage points) are more than twice those of any other model. Gemma 3 4B is the least sensitive by movement (9.8, with 8.10 shared majors), and Qwen 3.5 4B (17.3), GPT-OSS 20B (14.8), and Llama 3.1 8B (13.4) lie in between.

The two kinds of measure rank the models differently. Llama 3.1 8B has the smallest STEM shift of all (0.7 points), yet replaces about 2.3 of ten majors, whereas Gemma replaces fewer majors (about 1.9) but shifts STEM exposure by 2.3 points. Llama's cue sensitivity thus consists largely of substitutions within fields, which a STEM-exposure summary almost entirely misses. The rank-exposure encoding of Section 3.2 is designed to register this pattern. No directional STEM test met the audit's multiplicity-adjusted flagging rule for any model. Before mitigation, then, the dominant form of cue sensitivity is instability in which majors are recommended and where they appear, rather than a consistent steering of one group away from STEM.

The mitigation panel shows the same instability on the gender and race contrasts. Averaged across models, matched lists from original generation share 7.47 of ten majors, so changing only the demographic description replaces about 2.5 recommended majors. Sexual orientation and socioeconomic status, which the audit also covers, are not among the mitigation contrasts, so the results below concern gender and race.

5.4 Joint Effectiveness and Component Contributions from Ablation Analyses

Table 5 reports the common-policy comparison.

Table 5: Common-policy comparison on held-out profiles, averaged equally across the five models. Entries are score points (the recorded statistic multiplied by 100), except the STEM gap, in percentage points. The response index is $100P/P_0$, with absolute responses P from Appendix B.3; values above 100 indicate more differentiation among academic profiles than original generation, so the 101.2 of conventional reward reflects $P = 0.333$ against $P_0 = 0.329$. Bold marks the best value in each discrepancy and fit column; only robust DACA attains it in all five. These rows apply the same policy to every model and differ from the model-specific rules of Section 5.6.

	Cue discrepancy (lower is better)			Usefulness			
Policy	Movement $100M$	STEM gap (pp)	Exposure $100E$	Fit $100U \uparrow$	Lowest-cue fit $\uparrow$	Response index	Source distance
Original generation	15.6	4.51	15.5	63.5	61.0	100.0	0.0
Cue omission	**0.0**	**0.00**	**0.0**	63.5	63.5	91.6	15.9
Conventional reward	2.0	0.04	3.9	86.3	86.3	101.2	45.5
Aligned reward	2.0	**0.00**	3.9	**86.4**	**86.4**	100.9	45.3
Robust DACA	**0.0**	**0.00**	**0.0**	**86.4**	**86.4**	100.0	45.3

Robust DACA removes cue discrepancies and improves fit. Robust DACA gives matched variants identical lists: movement falls from 15.6 to zero, the STEM gap from 4.51 points to zero, and the exposure discrepancy from 15.5 to zero, as Corollary 4.4 predicts for a policy that meets its conditions. This consistency is accompanied by gains in academic fit. Average academic fit rises by

22.9 points, from 63.5 to 86.4, and fit for the least favorably treated cue rises by 25.4 points, from 61.0 to 86.4. The gain is therefore largest for the cue variant that was worst served. The response index stays at 100.0: the policy differentiates among academic profiles as much as the original does.

Cue consistency and recommendation quality. Cue omission also yields identical lists, as predicted, because deleting the cue makes matched prompts the same. It leaves fit at 63.5, however, and lowers the response index to 91.6. Relative to omission, robust DACA adds 22.9 fit points and 8.4 index points. Its larger source distance, 45.3 against 15.9, reflects the revision needed to improve the recommendations rather than a cost of consistency. Conventional reward, which does not enforce consistency, revises the original lists by a similar 45.5.

Reward modeling supplies most of the gain. Conventional reward reranking raises fit from 63.5 to 86.3 and reduces movement from 15.6 to 2.0, an 87.2% reduction, with a response index of 101.2. Reranking by a fit criterion learned from a rubric that ignores the cue thus removes most cue sensitivity while improving fit. Adding the matched-cue penalty (aligned reward) closes the remaining STEM gap of 0.04 points. Development diagnostics report a nearly zero matched-score residual, consistent with Lemma 4.2 when the separation loss is small. Yet movement remains at 2.0 and the exposure discrepancy at 3.9.

Complete-output alignment removes the residual. Aligned reward therefore has stable scores and no aggregate STEM gap, yet its lists still differ across cues. This matches the second prediction. By Theorem 4.3, the residual must enter through the candidate pools, the selection rule, or score differences near the top-ten boundary that an average score guarantee does not exclude. It is exactly the substitution pattern of Section 3.2. Adding a common pool and cue-marginalized selection closes all three routes and removes the residual movement of 2.0 and exposure discrepancy of 3.9, at no cost in fit (86.4 in both) or in revision (45.3 in both). The experiment evaluates the two components jointly and does not separate their individual contributions. Across Table 5, robust DACA is the only policy that attains the best value on every discrepancy and fit measure while preserving the original responsiveness.

5.5 Consistency Across Comparison Types and Cue Forms

Table 6: Cue discrepancy by comparison type on held-out profiles, averaged equally across models. Every contrast holds the academic profile fixed; "race within gender" changes race at a fixed stated gender, as in Asian woman versus Black woman. Movement and STEM gap are separate outcomes with different units.

		Movement (score points)		STEM gap (pp)	
Comparison type	Contrasts	Original	Robust DACA	Original	Robust DACA
Gender	4	14.32	0.00	4.14	0.00
Race	4	16.30	0.00	5.12	0.00
Race within gender	2	16.20	0.00	4.20	0.00

Table 6 separates the three comparison types. Under original generation, race contrasts produce more movement and larger STEM gaps than gender contrasts (16.30 against 14.32 points, and 5.12 against 4.14 percentage points). Race comparisons within a stated gender behave like race contrasts (16.20 and 4.20). In terms of options, original generation replaces about 2.4 majors under gender contrasts and 2.6 under the two race-comparison types. Robust DACA removes both discrepancies in every category, as its construction guarantees for every contrast in $\mathcal{C}_A$. Complete-list agreement holds within each comparison type as well as in the pooled mean.

The signed differences show why several cue forms must be audited (Appendix B.4). Under original generation, the woman-minus-man STEM-exposure difference is -1.45 points with explicit labels but $+0.07$ with pronouns, and the nonbinary-minus-man difference is -2.95 with labels but $+0.77$ with pronouns. The Black-minus-White difference is -1.41 with labels and -2.30 with paraphrase, and the Asian-minus-Black difference is -0.59 among profiles described as women but $+0.91$ among those described as men. These are descriptive panel results without pair-specific tests. Their small magnitudes and changing signs agree with the audit's finding that cue sensitivity takes the form of unstable option choices rather than a consistent directional penalty. They also show that removing an explicit label does not settle the question: the same attribute conveyed through pronouns or paraphrase produces different effects. Every signed difference is zero under robust DACA.

5.6 Model-Specific Selection and Held-Out Acceptance

For each model, a recommendation rule is selected on validation profiles and frozen. It is then tested on held-out profiles against prespecified requirements: upper limits on mean and worst-case movement, exposure shift, and source distance, and lower limits on mean and lowest-cue fit, profile response, and candidate coverage. Upper limits apply to upper confidence bounds and lower limits to lower bounds, and every requirement must pass (Appendix C). Table 7 summarizes the outcome for each model alongside its original cue sensitivity.

Selection on validation profiles. Robust DACA was selected for GPT-OSS, Llama, and Gemma. For the other two models, a safeguard overruled it. DeepSeek's robust rule reached zero movement and 95.8 fit points but revised the original recommendations beyond the 50-point source-distance ceiling, so the smaller revision of cue omission was selected. Qwen's robust rule also reached zero movement, with 88.6 fit points, but its absolute profile response was 0.68 points against a 5-point floor. Omission also fell below the floor, at 4.66, so the original policy was retained without certification. The Qwen case illustrates the utility floor of Section 4.2. An update that is consistent and scores well on fit but barely differentiates among academic profiles is rejected.

Held-out outcomes. On held-out profiles, four of the five selected rules meet every cue-stability requirement; the exception is Qwen's retained original policy, whose mean movement is 14.3 points. Llama's robust rule meets every requirement, with a mean fit of 83.1 and a lower bound of 77.0.

Table 7: Cue sensitivity and outcomes by model, ordered by original sensitivity. The first three columns come from the separate motivating audit of original generation (Appendix B.1), which uses different profiles and additional cue dimensions. On the held-out panel, robust DACA has zero movement and zero STEM gap for every model. The last two columns give the rule selected on validation profiles and any requirement it did not meet on held-out profiles (Appendix C).

	Original sensitivity (audit)					
Model	Shared majors	Move-ment	STEM shift	Robust DACA fit	Selected rule	Held-out requirement not met
DeepSeek-R1 8B	4.66	36.2	6.1	95.8[a]	Cue omission	Mean and lowest-cue fit
Qwen 3.5 4B	7.09	17.3	2.7	88.6[a]	Original (fallback)	Cue stability; mean and lowest-cue fit
GPT-OSS 20B	7.72	14.8	2.7	80.0	Robust DACA	Source distance (upper bound)
Llama 3.1 8B	7.69	13.4	0.7	83.1	Robust DACA	None: all requirements met
Gemma 3 4B	8.10	9.8	2.3	90.0	Robust DACA	Source distance (upper bound)

[a] Its fit is on the 0–100 scale and comes from held-out profiles, except for values marked [a], which come from validation profiles because a different rule was selected for those models.

Gemma's and GPT-OSS's robust rules meet every requirement on point estimates, including source distances of 47.3 and 45.5 under the 50-point ceiling, but their upper bounds of 55.3 and 53.8 exceed it. Compliance is therefore not established, which reflects insufficient precision rather than an observed violation. DeepSeek's omission rule eliminates movement but misses both fit floors, with a mean fit of 53.2 and a lower bound of 42.0.

What the per-model results show. Robust DACA removes cue discrepancies for every model, including DeepSeek, whose original sensitivity was the greatest. It delivers high fit for each: between 80.0 and 90.0 points on held-out profiles where it was selected, and 95.8 and 88.6 points on validation profiles for DeepSeek and Qwen. Where it was not adopted, a safeguard rather than a failure of alignment was responsible. For DeepSeek, whose omission rule reaches only 53.2 fit points against the 63.5-point average of original generation, the improvement required a larger revision than the prespecified ceiling allowed; for Qwen, the aligned lists were insufficiently differentiated across profiles. The requirements are thus consequential and model-specific. A model whose original recommendations are weak may need a large revision to become both consistent and useful. A ceiling on revision should therefore be set with the original model's quality in mind.

6 Discussion and Conclusion

DACA treats debiasing as a decision-level problem: remove the prohibited effect of an audited cue on the recommendation a person receives, while preserving the differences that legitimate information justifies. It connects what the system learns to value with what it delivers. Together, the theory

and the experiment support four lessons that extend beyond college-major advice.

6.1 Key Insights

Parity in a summary is not parity in the decision. An aggregate measure can certify a system that still treats matched cases differently. With aligned reward, the STEM-exposure gap was zero and matched scores were nearly identical, yet the recommended majors and their ranks still differed across cues. Before mitigation, Llama 3.1 8B shifted STEM exposure by only 0.7 points while replacing about 2.3 of ten majors. Theorem 4.3 explains why such residuals should be expected. A cue that no longer moves scores can still reach the output through the candidates that are generated and the rule that selects among them. For group exposure in multi-component recommenders, end-to-end fairness has been found to be largely achievable by improving individual components [26]; our results show that this conclusion need not carry over to within-profile consistency. Because an audit certifies consistency only for the output features it measures, decision-level audits should compare the identities and ranks of the options delivered, not only their aggregate composition.

Consistency and quality reinforced each other. Debiasing is often framed as a trade-off between fairness and accuracy. Here, the two improved together. Reranking by a fit criterion learned from a cue-blind rubric removed 87.2% of cue-induced movement while raising academic fit from 63.5 to 86.3. Because this policy keeps each variant's own candidates and changes only how they are scored and selected, the result suggests that most of the original cue effect entered through the model's ranking rather than through the options it generated. Completing the alignment did not reduce academic fit or increase source distance. The least favorably treated cue variant gained the most, 25.4 points against 22.9 on average. Cue omission, by contrast, produced consistency without better fit and with lower responsiveness. As the introduction argued, specifying what makes a recommendation good is part of addressing bias, not a separate performance concern.

Consistency alone is insufficient for acceptance. Exact consistency can be obtained by deleting the cue or returning the same list to every student. The safeguards of C5 determine whether such a policy is acceptable, and they affected selection in this experiment. Qwen's robust update was consistent and scored well on fit, yet it was rejected because it barely differentiated among academic profiles, the profile-insensitive outcome that the utility floor of Section 4.2 is designed to exclude. Held-out testing also separated two kinds of shortfall. DeepSeek's omission rule violated the fit floor, whereas the robust rules for Gemma and GPT-OSS met every requirement on point estimates but could not certify the source-distance ceiling at the achieved precision. Acceptance thresholds and evaluation precision should therefore be planned together. Allowing a no-update outcome makes both binding conditions for adoption. This is consistent with the pre-deployment evaluation called for in proposals for AI governance [34].

Bias took different forms across models, so correction must be structural. The five models differed not only in how cue-sensitive they were but in how that sensitivity appeared. DeepSeek-R1 8B changed more than half of each list, and Llama 3.1 8B substituted majors within fields while barely moving STEM exposure. Signed STEM differences were small and changed sign with the wording of the cue, and no directional test in the audit met its adjusted threshold for any model. A correction tuned to a presumed direction of bias, such as raising one group's STEM exposure, would therefore target the wrong object. DACA instead closes the cue's route into every stage of the pipeline, and its robust rule eliminated cue discrepancies for all five models, although which rule was adopted still depended on model-specific requirements. The DeepSeek case suggests that a model whose original recommendations are weak may need a large revision to become both consistent and useful. Limits on revision should therefore be set with the original model's quality in mind.

6.2 Scope and Limitations

The framework applies wherever permissible inputs, audited cue contrasts, a common output encoding, and an external standard of quality can be specified in advance. This decision-specific specification can be adapted to other settings. The college-major rubric and its thresholds are specific to this application. Applications in hiring, clinical decision support, or financial advice would supply their own quality standards and an explicit account of which attributes may play which roles, including attributes with both a permitted and a prohibited role (Section 3.1).

The evidence has clear limits. It concerns synthetic academic profiles, five open-weight models, ranked recommendations from a fixed catalog, and rubric-based utility. It demonstrates mitigation under that evaluation standard rather than effects on students' eventual choices or outcomes. The mitigation contrasts cover gender and race, whereas the motivating audit also covered sexual orientation and socioeconomic status. Consistency on the prespecified contrasts is not a claim about every proxy for identity, such as names or dialect [8, 2]. The ablation evaluates the common pool and cue-marginalized selection jointly, so their separate contributions remain to be identified. Finally, deploying robust DACA requires constructing counterfactual variants of each request, which presupposes that the audited cue can be located and substituted in the user's input and multiplies generation cost by the number of cue values and replicates.

These limits define the next steps: broader and real-world profile samples; blinded expert judgments or validated outcomes as the external standard of quality; validated encodings for free-text recommendations; audits that extend to proxies such as names and dialect; generators trained toward common support where model weights are available; and amortized pooling to reduce the cost of inference.

6.3 Conclusion

DACA makes the delivered recommendation the common target of reward modeling, counterfactual alignment, and independent validation. Its theory shows that a cue can reach a recommendation

through candidate generation, scoring, and selection, and that exact alignment is guaranteed when all three are controlled but not, in general, when any one is left open. Its experiment shows that such control can coexist with better advice: robust DACA removed cue discrepancies for all five models while raising average academic fit from 63.5 to 86.4 and preserving responsiveness to academic differences. The contribution is a procedure for pursuing invariance together with useful personalization and explicit criteria for accepting the result. Advice should differ when legitimate decision evidence differs, and not when the same case is presented under a different demographic description.

Supplementary information

Supplementary Material contains Appendices A–C: proofs and technical details, supporting empirical evidence, and recommendation-policy selection and held-out testing.

References

[1] T. Zack, E. Lehman, M. Suzgun, et al., "Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: A model evaluation study," *The Lancet Digital Health*, vol. 6, no. 1, pp. e12–e22, 2024. https://doi.org/10.1016/S2589-7500(23)00225-X.

[2] H. Nghiem, J. Prindle, J. Zhao, and H. Daumé III, "You gotta be a doctor, Lin: An investigation of name-based bias of large language models in employment recommendations," in *Proceedings of EMNLP*, pp. 7268–7287, 2024. https://doi.org/10.18653/v1/2024.emnlp-main.413.

[3] 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.

[4] M. S. Khan, H. Umer, H. Mahmud, and S. Rothenberg, "When personalization becomes bias: Structural and discursive religious framing in AI-generated financial advice," arXiv:2608.16909v1, 2026. https://arxiv.org/abs/2608.16909v1.

[5] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019. https://doi.org/10.1126/science.aax2342.

[6] A. Lambrecht and C. E. Tucker, "Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads," *Management Science*, vol. 65, no. 7, pp. 2966–2981, 2019. https://doi.org/10.1287/mnsc.2018.3093.

[7] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, 2017. https://doi.org/10.1126/science.aal4230.

[8] V. Hofmann, P. R. Kalluri, D. Jurafsky, and S. King, "AI generates covertly racist decisions about people based on their dialect," *Nature*, vol. 633, pp. 147–154, 2024. https://doi.org/10.1038/s41586-024-07856-5.

[9] M. Leyva-Vázquez and F. Smarandache, "Institutional prestige as geographic bias in large language models: Evidence from three factorial experiments with bootstrap confidence intervals," arXiv:2608.18107v1, 2026. https://arxiv.org/abs/2608.18107v1.

[10] A. Singh and T. Joachims, "Fairness of exposure in rankings," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2219–2228, 2018. https://doi.org/10.1145/3219819.3220088.

[11] M. Bertrand and S. Mullainathan, "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination," *American Economic Review*, vol. 94, no. 4, pp. 991–1013, 2004. https://doi.org/10.1257/0002828042002561.

[12] K. Vafa, S. Athey, and D. M. Blei, "Estimating wage disparities using foundation models," *Proceedings of the National Academy of Sciences*, vol. 122, no. 22, e2427298122, 2025. https://doi.org/10.1073/pnas.2427298122.

[13] I. Ng, P. Blöbaum, S. Bhandari, K. Zhang, and S. Kasiviswanathan, "Debiasing reward models by representation learning with guarantees," arXiv:2510.23751v1, 2025. https://doi.org/10.48550/arXiv.2510.23751.

[14] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth, "The reusable holdout: Preserving validity in adaptive data analysis," *Science*, vol. 349, no. 6248, pp. 636–638, 2015. https://doi.org/10.1126/science.aaa9375.

[15] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems*, vol. 29, pp. 3315–3323, 2016.

[16] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226, 2012. https://doi.org/10.1145/2090236.2090255.

[17] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[18] R. Nabi and I. Shpitser, "Fair inference on outcomes," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018. https://doi.org/10.1609/aaai.v32i1.11553.

[19] J. Zhang, K. Bao, Y. Zhang, W. Wang, F. Feng, and X. He, "Is ChatGPT fair for recommendation? Evaluating fairness in large language model recommendation," in *Proceedings of the 17th ACM Conference on Recommender Systems*, 2023. https://doi.org/10.1145/3604915.3608860.

[20] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *Proceedings of the 30th International Conference on Machine Learning*, PMLR, vol. 28, no. 3, pp. 325–333, 2013. https://proceedings.mlr.press/v28/zemel13.html.

[21] D. Madras, E. Creager, T. Pitassi, and R. Zemel, "Learning adversarially fair and transferable representations," in *Proceedings of the 35th International Conference on Machine Learning*, PMLR, vol. 80, pp. 3384–3393, 2018.

[22] S. Garg, V. Perot, N. Limtiaco, A. Taly, E. H. Chi, and A. Beutel, "Counterfactual fairness in text classification through robustness," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 219–226, 2019. https://doi.org/10.1145/3306618.3317950.

[23] W. Hua, Y. Ge, S. Xu, J. Ji, Z. Li, and Y. Zhang, "UP5: Unbiased foundation model for fairness-aware recommendation," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1899–1912, 2024. https://doi.org/10.18653/v1/2024.eacl-long.114.

[24] C. Dwork and C. Ilvento, "Fairness under composition," in *10th Innovations in Theoretical Computer Science Conference*, LIPIcs, vol. 124, pp. 33:1–33:20, 2019. https://doi.org/10.4230/LIPIcs.ITCS.2019.33.

[25] C. Dwork, C. Ilvento, and M. Jagadeesan, "Individual fairness in pipelines," in *1st Symposium on Foundations of Responsible Computing*, LIPIcs, vol. 156, pp. 7:1–7:22, 2020. https://doi.org/10.4230/LIPIcs.FORC.2020.7.

[26] X. Wang, N. Thain, A. Sinha, F. Prost, E. H. Chi, J. Chen, and A. Beutel, "Practical compositional fairness: Understanding fairness in multi-component recommender systems," in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pp. 436–444, 2021. https://doi.org/10.1145/3437963.3441732.

[27] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

[28] I. O. Gallegos, R. A. Rossi, J. Barrow, et al., "Bias and fairness in large language models: A survey," *Computational Linguistics*, vol. 50, no. 3, pp. 1097–1179, 2024. https://doi.org/10.1162/coli_a_00524.

[29] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A reductions approach to fair classification," in *Proceedings of the 35th International Conference on Machine Learning*, PMLR, vol. 80, pp. 60–69, 2018. https://proceedings.mlr.press/v80/agarwal18a.html.

[30] H. Gonen and Y. Goldberg, "Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them," in *Proceedings of NAACL-HLT*, pp. 609–614, 2019. https://doi.org/10.18653/v1/N19-1061.

[31] N. Meade, E. Poole-Dayan, and S. Reddy, "An empirical survey of the effectiveness of debiasing techniques for pre-trained language models," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 1878–1898, 2022. https://doi.org/10.18653/v1/2022.acl-long.132.

[32] L. Ouyang, J. Wu, X. Jiang, et al., "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems*, vol. 35, pp. 27730–27744, 2022. https://doi.org/10.52202/068431-2011.

[33] 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.

[34] Y. Bengio, G. Hinton, A. Yao, et al., "Managing extreme AI risks amid rapid progress," *Science*, vol. 384, no. 6698, pp. 842–845, 2024. https://doi.org/10.1126/science.adn0117.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- itmappendix.pdf