Mitigating Label Bias with Interpretable Rubric Embeddings

Calvin Isley
Harvard Kennedy School
Harvard University
Cambridge, MA 02139
cisley@g.harvard.edu

Johann D. Gaebler
Department of Statistics
Harvard University
Cambridge, MA 02139
jgaebler@fas.harvard.edu

Sharad Goel
Harvard Kennedy School
Harvard University
Cambridge, MA 02139
sgoel@hks.harvard.edu

Abstract

Statistical decision algorithms are increasingly deployed in domains where ground-truth labels are hard to obtain, such as hiring, university admissions, and content moderation. In these settings, models are typically trained on historical human evaluations—for example, using past hiring decisions as a proxy for true applicant quality. However, if past evaluations unjustly favor certain groups, models trained on these labels may inherit those biases. To address this problem, we propose basing predictions on rubric embeddings, a representation framework that replaces standard black-box embeddings with features derived from expert-defined criteria that align with the underlying construct of interest. By anchoring predictions to semantically meaningful dimensions, this approach guards against biased proxy signals. We provide both theoretical and empirical evidence that rubric embeddings mitigate label bias under plausible conditions. Empirically, we evaluate our method on a novel dataset of applications to a large master's program. We find that models trained on rubric embeddings reduce group disparities while improving measures of cohort quality. Our results suggest that basing predictions on interpretable, domain-grounded representations offers a practical approach to learning in the presence of biased labels.

1 Introduction

In many supervised learning problems, models are trained to predict labels that are treated as ground truth for the quantity of interest. This includes, for example, predicting object categories from images, transcripts from speech, and disease status from medical data. In many domains, however, ground-truth labels do not exist or are prohibitively difficult to obtain. Often the target quantity is a latent construct that is only imperfectly defined and can only be observed for a non-representative subset of individuals. For instance, in hiring, one may wish to select the candidate most likely to succeed in a role, but "success" is inherently ambiguous and is typically only observed for candidates who were hired in the past.

In the absence of ground-truth labels, a common empirical strategy is to predict historical human evaluations, such as hiring decisions or evaluator ratings of past applicants. But these proxy labels may be biased, reflecting implicit and unjustified preferences for certain groups. As a result, models trained on these proxy labels can lead to suboptimal and inequitable decisions. Indeed, in a widely reported case, an internal hiring model at Amazon penalized resumes containing terms such as "women's" (e.g., as in "women's chess club captain"), ostensibly reflecting biases in the historical hiring data [1].

In this paper, we show how models trained on proxy labels can inherit and amplify bias. Using a novel dataset of applications to a master's program, we show that the common approach of training

Preprint.

models on black-box text embeddings can reproduce biases present in historical evaluations. In particular, we find that such embeddings encode sensitive attributes like gender, even when explicit markers are removed, and that models trained on these representations are, under certain conditions, guaranteed to replicate biases in the data. We provide both empirical and theoretical evidence for this phenomenon, drawing on the framework of correspondence experiments from the social sciences.

We then analyze three past approaches to mitigate this problem—orthogonalization, redaction, and marginalization—and discuss their limitations. In brief, orthogonalization induces demographic parity, which can inadvertently penalize qualified applicants from certain groups. Redaction, while sound in theory, can fail in practice, as residual signals of gender often remain in the embeddings. Finally, marginalization ensures that certain types of biases are removed, but can introduce other forms.

Given these limitations, we propose instead basing predictions on rubric embeddings, representations constructed from expert-defined criteria aligned with the outcome of interest. In the case of university admissions, we construct hundreds of rubric items that encode, for example, past coursework and grades, work experience, and letter evaluations. We then use LLMs to score applications along the rubric dimensions to create a semantically grounded representation of the data. We show that basing predictions on rubric embeddings addresses the central limitations of past methods, yielding cohorts of students that are both highly qualified and demographically diverse. Our results suggest that rubric embeddings provide a practicalstrategy for mitigating label bias in a wide range of decision-making contexts, addressing a long-standing challenge in the equitable design of algorithms [2]. We conclude by connecting our work to two common understandings of discrimination in the law—disparate treatment and disparate impact—and discussing some of the limitations of our approach.

Related Work. Our work connects three strands of research that have developed largely in parallel: (1) algorithmic fairness and, specifically, the problem of label bias; (2) emerging work on concept bottleneck models and rubric representations, which considers how best to construct and leverage interpretable, semantically meaningful features; and (3) the design of correspondence experiments, also known as audit studies, popular for measuring discrimination in the social sciences. We discuss the first two areas below, and then consider correspondence experiments in detail in Section 2.

An extensive literature in algorithmic fairness develops techniques for identifying and mitigating biases in algorithmic systems [see, e.g., 2–9]. We focus on *label bias*, which arises when models are trained on proxies that systematically diverge from the underlying outcomes of interest. For example, an algorithm that predicts future healthcare expenditures as a proxy for future patient needs can induce racial disparities, since White patients, on average, spend more than equally sick Black patients [10]. Ideally, one would acquire more accurate labels, but doing so is often expensive or otherwise infeasible. To avoid collecting new data, past work has considered reweighting the existing, biased data under structural assumptions about the labeling process [11], or adjusting the loss to account for group-dependent label noise [12]. These approaches aim to correct bias at the level of the training objective, but rely on strong assumptions about how bias enters the labeling process, which may not hold in practice. Zanger-Tishler et al. [13] show that, in the presence of label bias, selectively removing features can improve performance on the true, unobserved label. Theirs is an important theoretical insight, but they do not offer a constructive method for selecting the features to remove. Building on that work, we show that using rubric embeddings provides a practical, constructive method for selecting which features to retain, effectively operationalizing this idea of feature removal; we analytically relate our approach to their observation.

Concept bottleneck models (CBMs) structure predictions by factoring them through a set of interpretable, human-defined features [14]. Originally developed for image data, CBMs have been extended to text, where LLMs are used to score documents on predefined or generated concept sets, often achieving performance competitive with black-box approaches [15–17]. Rubric embeddings (also known as rubric-derived representations) are a closely related approach that uses LLMs to extract structured features from text, and have likewise shown competitive or superior performance to black-box embeddings in low-data regimes [18–20]. Although CBMs and rubric-derived representations have developed largely in parallel, both share the central idea of substituting a small, semantically meaningful feature set for a generic embedding. Most directly related to our own work, Ludan et al. [17] and Demirel et al. [20] both use LLMs to generate a collection of interpretable, semantically relevant features. This literature has focused primarily on interpretability, sample efficiency, and

robustness to distribution shift [21, 22]. However, to our knowledge, the potential of concept-based representations to mitigate label bias has not been studied.

2 Label Bias and Black-Box Embeddings: A Case Study in Admissions

We begin by illustrating the problem of label bias and discussing its connection to black-box embeddings. To do so, we imagine designing an algorithm to evaluate applicants to a graduate degree program. We base our empirical analysis on datasets derived from real applications to a large master's program in public policy, where we synthetically inject increasing degrees of bias into the real, observed admissions decisions. This approach allows us to study label bias across a range of settings. In the process, we build on and formalize the notion of correspondence studies, a widely used method for measuring discrimination in the social sciences.

We specifically start with 1,112 applications submitted during the 2025–2026 application cycle. Each application contains structured information, such as basic demographics and standardized test scores, along with four kinds of unstructured materials: a resume of professional and volunteer experience; transcripts from all previous undergraduate and graduate institutions; three letters of recommendation; and short essays describing applicants' backgrounds, reasons for applying, and future plans. Admissions officers evaluate each application through a structured process resulting in a 4–20 point score. These scores determine admissions decisions and are distributed approximately normally, with most applicants receiving between 10 and 17 points (Fig. A1).

For the purposes of our stylized empirical analysis, we treat these numeric scores as ground-truth labels. While these observed scores may themselves contain noise or bias, they serve as a fixed reference point for evaluating the relative performance of different modeling approaches. We then explicitly introduce label bias by adding a group-specific noise term to the scores. Specifically, given the actual score Y_i for the i-th applicant, we define the proxy score

$$Y_i' \stackrel{\text{def}}{=} \begin{cases} Y_i + Z_i & G_i = m, \\ Y_i - Z_i & G_i = f, \end{cases} \qquad Z_i \sim \mathcal{N}(b, 1), \qquad Z_i \perp\!\!\!\perp X_i, G_i, \tag{1}$$

where $G_i \in \{f, m\}$ denotes whether the applicant is male or female and b gives the size of the advantage for male applicants. We vary b from 0, representing no advantage, to 2.5, representing a large advantage for male applicants.

Throughout, we fit models on the proxy label Y', and then consider both the "true" quality (as indicated by Y) and demographic composition of the cohort of top-ranked applicants under the learned model. To fit these models, we must construct a representation of the unstructured data; one common approach is to embed the data using general-purpose black-box text embeddings [23–25].

2.1 Theoretical analysis

We quantify the bias of the fitted models by drawing on the extensive literature on correspondence experiments [26]. Correspondence experiments, also known as "audit studies," measure the extent to which decisions in hiring, admissions, and related settings vary for individuals who are (nearly) identical except for some protected characteristic [27, 28]. In practice, researchers experimentally manipulate materials to alter decision-makers' perceptions of group membership. In one of the first such studies, Bertrand and Mullainathan [29] sent out job applications that were identical in every respect but the name of the applicant. Those with names suggesting they were Black (e.g., 'Lakisha' or 'Jamal') received significantly fewer callbacks than those with names suggesting they were White (e.g., 'Emily' or 'Greg')—providing evidence of racial bias.

To formalize this approach to measuring bias, let $T(x, g) : \mathcal{X} \times \{m, f\} \to \mathcal{X}$ denote a correspondence experiment *manipulation*: a (possibly stochastic) transformation of an applicant's features $X \in \mathcal{X}$ with $T(x, m)$ and $T(x, f)$ representing the "male-presenting" and "female-presenting" versions of $x \in \mathcal{X}$, respectively.[1] For example, one such transformation might change the (raw, unstructured) application materials so that the applicant's listed name and the pronouns their letter writers use to refer to them accord with the target gender. Now, given an algorithm $h(x) : \mathcal{X} \to \mathbb{R}$, we define the bias of h under the correspondence experiment determined by T as follows.

[1] For notational conventions, full details on regularity conditions, and proofs, see Appendix A.

Definition 1 ($\text{bias}_T(h)$)**.** We define the *bias* with respect to $T(x, g)$ as

$$\text{bias}_T(h) \stackrel{\text{def}}{=} \mathbb{E}[h(T(X, m)) - h(T(X, f))]. \tag{2}$$

Def. 1 captures the average extent to which predictions for an applicant differ between their male- and female-presenting features. With this framework, we can now explicitly derive the bias of models trained with a broad range of biased labels, including those defined in Eq. (1). (Proofs of Prop. 2 and other results are given in Appendix A.)

Proposition 2. *Let $Y, B \in \mathbb{R}$ be random variables, and let $G \in \{m, f\}$, $X \in \mathcal{X}$, and $T(x, g) : \mathcal{X} \times \{m, f\} \to \mathcal{X}$ be as above. Let $r(x) : \mathcal{X} \to \mathcal{R}$ and define $R \stackrel{\text{def}}{=} r(X)$. Suppose $B \perp\!\!\!\perp X \mid G$. Then*

$$\text{bias}_T(\nu) = \text{bias}_T(\mu) + (\mathbb{E}[B \mid G = m] - \mathbb{E}[B \mid G = f]) \cdot \mathbb{E}[\pi(T(X, m)) - \pi(T(X, f))] \tag{3}$$

where $\nu(x) \stackrel{\text{def}}{=} \mathbb{E}[Y + B \mid R = r(x)]$, $\mu(x) \stackrel{\text{def}}{=} \mathbb{E}[Y \mid R = r(x)]$, and $\pi(x) \stackrel{\text{def}}{=} \Pr(G = m \mid R = r(x))$.

Prop. 2 considers the bias of an algorithm $\nu(x) = \mathbb{E}[Y + B \mid R = r(x)]$ that perfectly estimates a biased label $Y + B$ given the available information in some representation R of an application, such as a black-box text embedding. In it, we assume that B perturbs the ground truth label Y in a way that only depends on group membership G, as holds for our synthetic datasets described in Eq. (1). In this case, the proposition decomposes $\text{bias}_T(\nu)$ into three terms: (1) $\text{bias}_T(\mu)$, the baseline bias in the algorithm μ present when predicting the true label Y; (2) $\mathbb{E}[B \mid G = m] - \mathbb{E}[B \mid G = f]$, the difference in the average perturbation B across groups, a measure of how much a group is advantaged; and (3) $\mathbb{E}[\pi(T(X, m)) - \pi(T(X, f))]$, the difference in the probability that male- and female-presenting features belong to a male applicant. By Eq. (1),

$$\mathbb{E}[B \mid G = m] = \mathbb{E}[Z \mid G = m] = b, \qquad \mathbb{E}[B \mid G = f] = \mathbb{E}[-Z \mid G = f] = -b.$$

Consequently, for our stylized example,

$$\text{bias}_T(\nu) = \text{bias}_T(\mu) + 2b \cdot \mathbb{E}[\pi(T(X, m)) - \pi(T(X, f))]. \tag{4}$$

Critically, $\text{bias}_T(\nu)$ depends on the extent to which applications can be manipulated to look more male or more female. In other words, to the extent that R implicitly encodes group membership, it is more likely to inherit any biases present in the labels. Our example assumes a particular form of label bias to facilitate analysis, but the core qualitative insight extends beyond this specific formulation, a pattern we explore empirically in the subsequent sections.

2.2 Empirical analysis

We next empirically examine how much gendered information is implicitly encoded in standard text embeddings, and analyze its consequences for cohort selection. To convert the raw application materials to vector representations, we first extract their contents with Microsoft Document Intelligence [30]. We then separately embed each of the four document types (resumes, essays, letters, and transcripts) using OpenAI's `text-embedding-3-large` model [31]. Given the limited number of applications, we use only the first 250 dimensions of the embeddings for each category [32]. We additionally include some structured data, but we omit gender and other explicit demographic features because deployed admissions models likely cannot legally include them [33]. This process results in an approximately 1,000-dimensional embedding.

We next use these vector representations to predict the biased labels Y', for $b \in [0, 2.5]$. We specifically fit ridge regression models, which we denote by h, using the `glmnet` package [34]. To generate out-of-sample predictions, we use 10×10-fold nested cross-validation: For each of the 10 outer folds, we fit and tune a ridge model on the remaining nine folds using 10-fold cross-validation before predicting on the held-out fold. We then calibrate the resulting models using linear scaling on the training folds. We repeat this process 20 times, combining standard errors across iterations using Rubin's rules [35]. We fit models on a 4-core CPU, running for roughly 12 hours.

Following Gaebler et al. [26], we define $T(x, g)$ to account for many signals of gender beyond names that rich inputs like our application materials contain. We use language models to manipulate gendered cues about applicants, including: (1) third-person pronouns; (2) gendered titles (e.g., "Mr."

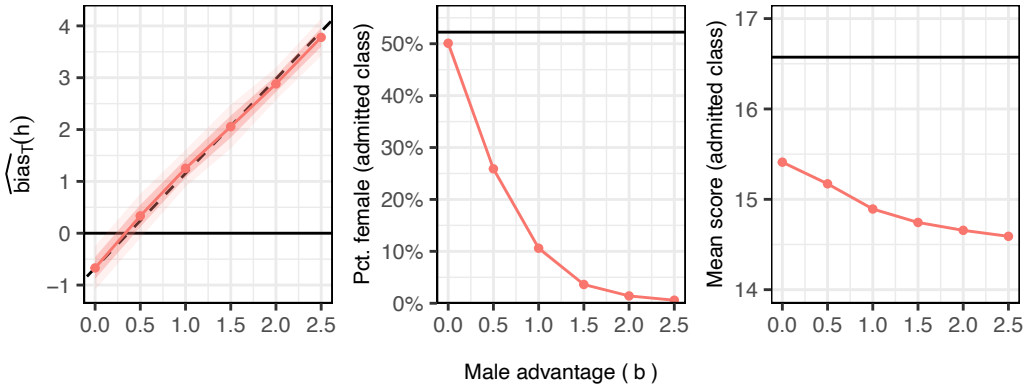

Figure 1: *Bias of regression models trained on proxy labels Y' and consequences for admitted classes. The x-axis in all panels denotes male advantage b in Eq.* (1)*. Dark and light shaded regions show pointwise 68% and 95% confidence intervals (not visible in center and right-hand panels). Solid black lines show zero bias (left) or corresponding values for the top 20% of students ranked by actual score Y (center, right).* Left: *Estimated bias $\widehat{\text{bias}}_T(h)$ of fitted models $h(x)$. The dashed line shows bias estimated according to Eq.* (3)*.* Middle: *Percentage of women in the admitted class under a top-20% policy.* Right: *Mean true score Y of the admitted class.*

or "Mrs."); (3) kinship terms (e.g., "son" or "daughter"); and (4) gender-marked occupations (e.g., "waitress") and gender-exclusive organizations (e.g., fraternities or sororities). Because applicants' home countries often feature prominently in their application materials, we replace their name with that of a randomly chosen historical applicant from the same country with the appropriate gender. Further details, including our validation process and prompt templates, appear in Appendix B.

By Eq. (4), $\text{bias}_T(h)$ equals the baseline bias in h plus twice the product of b and $\mathbb{E}[\pi(T(X,m)) - \pi(T(X,f))]$, where the latter term quantifies the extent to which our representations encode gender. In our data, we estimate this difference is $91.2\% \pm 0.6\%$, meaning that altering applications to indicate gender strongly impacts the embeddings, which in turn impacts the model's predictions of gender. Relatedly, ridge regression models predicting self-reported gender from the text embeddings achieve an out-of-sample AUC of $99.6\% \pm 0.1\%$, again demonstrating that the embeddings encode gender with near-perfect accuracy. While this empirical result is not surprising given how well standard black-box text embeddings encode information, it does suggest that the models will strongly inherit the bias injected into the labels.

Further, the baseline $\text{bias}_T(\mu)$ equals -0.67, meaning that the algorithm is biased *against* men when predicting the ground-truth label. We note that this quantity represents the bias of the *algorithm* and not the bias of the admissions *decisions*. For predictive models trained using a broader collection of features, including information about past student performance not well-captured by black-box embeddings, this bias effectively vanishes, suggesting that the actual admissions decisions are not in fact biased; see Fig. 2.

The left panel of Fig. 1 shows that our theoretical estimates of $\text{bias}_T(h)$ (dashed line) almost perfectly match the empirical estimates from the data. The middle and right panels illustrate the consequences for admissions. We consider top-k admissions policies that rank applicants by predicted score and admit the top 20%, a hypothetical admissions rate in line with the admissions policies of competitive graduate programs. As the advantage given to male applicants grows, the proportion of women admitted falls precipitously (middle panel), dropping to effectively zero when $b = 2.5$. The quality of the admitted class, measured by applicants' "ground-truth" scores Y, declines in parallel (right panel), as biased predictions substitute less qualified men for more qualified women.

3 Mitigating Label Bias

Our extended example above illustrates how label bias can propagate to predictions in models based on black-box text embeddings, reproducing the reported issues with Amazon's internal hiring model. Our theoretical analysis suggests this phenomenon stems from the ability of black-box embeddings

to capture a rich set of signals in the data, including those related to gender. While typically an asset, this property can also result in the models inheriting bias in the labels. As such, it seems reasonable to select a representation of the data that does not so strongly encode these gender signals. But there are better and worse ways to do so. We first analyze three common approaches in the algorithmic fairness literature—orthogonalization, redaction, and marginalization—and discuss their limitations. We then introduce rubric embeddings, showing that they effectively eliminate label bias in our setting while avoiding the shortcomings of other methods.

3.1 Orthogonalization, redaction, and marginalization

Orthogonalization [36, 37] involves selecting a representation $R = r(X)$ such that $R \perp\!\!\!\perp G$, meaning the representation is (approximately) independent of group membership. Orthogonalization and the closely related technique of adversarial debiasing [38, 39] are common recommendations in the literature for achieving fair downstream predictions when training on rich features that can closely approximate group membership. Both techniques proceed by removing information about a protected characteristic like gender from individuals' feature representations while otherwise leaving the remaining information intact, for instance by projecting latent representations of applicants onto subspaces orthogonal to their gender. The end goal of orthogonalization and related techniques is to eliminate gender from an individual's features in an information-theoretic sense.

This approach, however, comes with substantial drawbacks. In particular, independence from the protected characteristic G carries through to downstream decisions based on the latent representation R. For instance, if $d(r) : \mathcal{R} \to \{0, 1\}$ represents an admissions policy, the proportion of admitted students from different groups will necessarily equal the population proportion, i.e.,

$$\Pr(G = g \mid d(R) = 1) = \Pr(G = g).$$

(See Appendix A for proof.) But if one group of applicants is more qualified than another—for instance, if female applicants are, on average, more qualified than male applicants—enforcing such demographic parity can inadvertently penalize that more qualified group. While in practice orthogonalization may only partially remove group information, even approximate independence can induce substantial shifts toward demographic parity. Indeed, while the aim of orthogonalization is to improve equity, it might itself constitute illegal gender (or racial) "balancing" [33]; though cf. [40, 41].

Another natural approach to reducing bias is to redact signals of gender from an applicant's materials *before* embedding them. Formally, redaction represents a map $\rho : \mathcal{X} \to \mathcal{X}$ with the property that it masks the transformation T, i.e.,

$$\rho(T(X, g)) = \rho(X), \qquad g \in \{m, f\}. \tag{5}$$

In our applied admissions example, for instance, we operationalize $\rho(x)$ by creating alternate gender-neutral representations of applicants' materials in which we obscure signals of gender instead of swapping their gender presentation (e.g., replacing "he" with "they" instead of "her"); see Appendix B for full details.

Redaction is guaranteed to be unbiased with respect to T (see Prop. A2). But unlike orthogonalization, the elimination of gender information through redaction is narrowly tied to a particular manipulation, potentially limiting its effectiveness. To evaluate this possibility, we train models predicting applicants' genders using black-box embeddings of the redacted materials. Despite our redaction efforts, we find that these models still predict applicants' genders with high accuracy, achieving an AUC of $81.2\% \pm 1.4\%$. Even in the absence of strong gender signals—like names and pronouns—the predictive models we train can nevertheless ascertain individuals' gender based on subtle statistical regularities encoded by the black-box embeddings, a finding consistent with past work [e.g., 26, 42, 43]. As a result, predictive models trained on black-box embeddings of facially gender-neutral materials can nevertheless reproduce label bias in their predictions, a point we return to below.

Finally, marginalizing out gender or other protected characteristics is a well-explored technique in economics [e.g., 44, 45], as well as the causal fairness literature [e.g., 46]. Marginalized predictions replace an individual's actual prediction with a weighted sum of the predictions made when their actual features are replaced with features belonging to different groups. Formally, marginalization replaces an algorithm $h(x) : \mathcal{X} \to \mathbb{R}$ with the weighted sum

$$h_T(x) \stackrel{\text{def}}{=} w \cdot h(T(x, m)) + [1 - w] \cdot h(T(x, f)), \tag{6}$$

for some weight $w \in [0, 1]$.

Like redaction, this approach results in predictive algorithms with no bias with respect to T (see Prop. A2.). However, despite its intuitive appeal, marginalization can lead to poor outcomes in practice when implemented using black-box embeddings, which we empirically illustrate below.

3.2 Rubric Embeddings

Orthogonalization, redaction, and marginalization share a common methodological goal: surgically eliminating information tainted by gender or other protected characteristics while preserving as much as possible of the rich information present in black-box embeddings. *Rubric embeddings* represent the inverse approach. Rather than removing potentially problematic information, rubric embeddings consist of only information deemed substantively important for the prediction task. Conceptually, rubric embeddings restrict the hypothesis class to functions of semantically meaningful features, limiting the ability of models to exploit spurious correlations. In domains like admissions, hiring, and medicine, expert decision makers can often articulate what they are looking for—even if they cannot specify a precise decision rule based on those dimensions. Enumerating these dimensions using a rubric against which unstructured materials can be scored (e.g., by grading them with a language model) allows the construction of a feature representation anchored to the aspects of the materials of greatest substantive relevance.

To demonstrate this approach, we develop a collection of 396 interpretable measures admissions officers deem substantively important in application review. In collaboration with the admissions staff at our partner master's program, we identify 14 measures of essay quality rated on five-point scales, ranging from grammatical correctness and clarity to prompt adherence and substantive appropriateness; 8 measures summarizing applicants' work histories, including tenure in different sectors and measures of seniority and career trajectory; 39 measures of letters of recommendation, including endorsement strength both overall and along a variety of personal dimensions, references to specific program-relevant technical skills, and substantive appropriateness of the choice of recommender; and 335 distinct measures of previous academic performance, including majors and minors, graduation honors, rank and other background information about previous institutions, and GPA both overall and in specific courses (e.g., microeconomics, calculus II), and areas (e.g., advanced undergraduate mathematics and statistics). We omit applicant demographics from our rubric embedding. Building on the document extraction pipeline used to extract black-box embeddings, we use OpenAI's GPT-5 series of models [47] to score the full applicant pool along each of the 396 dimensions.

The rubric embeddings are based largely on objectively measurable factors and do not include explicit indicators of gender. As such, for our usual transformation T, we expect that $r(T(X, g)) = r(X)$, implying that models based on rubric embeddings should be unbiased with respect to T. Below we empirically check this and examine the demographic composition and quality of the applicants models using rubric embeddings rank highest. We further compare this rubric embedding approach to several of the approaches discussed above: (1) predictive models trained on black-box embeddings in Section 2 ("black-box embeddings"), (2) predictive models trained using both black-box and rubric embeddings ("kitchen sink"), (3) predictive models trained on black-box embeddings of redacted application materials ("redacted embeddings"), and (4) marginalized black-box embedding models ("marginalized"). We omit comparisons with orthogonalized models, given their theoretical limitations.

The results are shown in Fig. 2. Across levels of male advantage, we find virtually zero bias in rubric embedding models, as expected (left panel). (We omit marginalized and redacted embedding models, which are guaranteed to have negligible bias.) The center and right-hand panels of Fig. 2 illustrate that rubric embeddings likewise admit the most gender-balanced and highest quality classes, across the full range of male advantage. This pattern remains true even when measured according to alternative measures of cohort quality; see Fig. A3.

While generally more performant than the other bias mitigation techniques, marginalized models produce admissions policies that admit substantially fewer women and a much less academically prepared class than rubric embedding models. This pattern is a consequence of the technique's forced bias correction. As shown in Fig. 1, the black-box embedding models exhibit a slight bias against male candidates when $b = 0$. Marginalizing out gender penalizes the group that would otherwise have an advantage and *vice versa*. In this case, that means the model penalizes women, leading to fewer

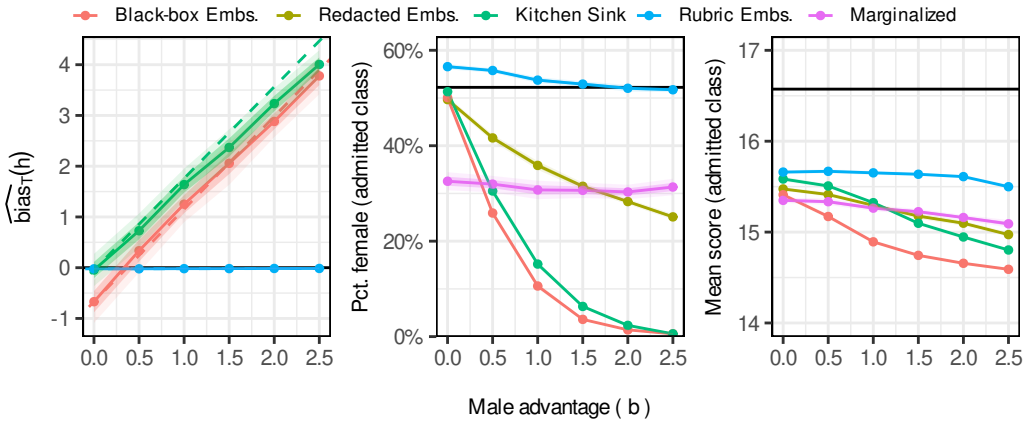

Figure 2: *Bias of models trained on proxy labels Y' under different bias mitigation techniques and consequences for admitted classes. The x-axis in all panels denotes male advantage b in Eq.* (1)*. Colors indicate the bias mitigation technique. Dark and light shaded regions show pointwise 68% and 95% confidence intervals. Solid black lines show zero bias (left) or corresponding values for the top 20% of students ranked by actual score Y (center, right).* Left: *Estimated bias $\widehat{\text{bias}}_T(h)$ of fitted models $h(x)$. Dashed line shows bias estimated according to Eq.* (3)*.* Middle: *Percentage of women in the admitted class under a top-20% policy.* Right: *Mean true score Y of the admitted class.*

women among the top-ranked applicants. Importantly, though, the black-box model's bias against men does not appear to reflect actual bias in the ground-truth labels Y. Indeed, the kitchen-sink models exhibit no such bias, suggesting that the bias in the black-box models results from their particular representation. Consequently, while the marginalized models attempt to "correct" apparent bias in the black-box models, they ultimately unjustly penalize female applicants.

Why do rubric embeddings perform so well relative to other techniques? Zanger-Tishler et al. [13] show that unlike when predicting ground-truth labels, including additional covariates can reduce prediction accuracy in the presence of label bias. In particular, they show that the relative predictive accuracy of models using a reduced and full set of features is controlled by three terms: (1) how accurately models trained on the reduced feature set can predict the proxy label, (2) how accurately models trained on the full feature set can predict the proxy label, and (3) how correlated the ground truth label is with the predictions using the full feature set, conditional on the reduced features. Since the models trained on the full set of covariates in expectation will predict the proxy label weakly better than models trained on the reduced set, the operative term is the correlation. The DAG shown in the left-hand panel of Fig. 3 approximately describes our admissions setting, where admissions officers attempt to score applicants based essentially entirely on the information captured in our rubric embeddings. In this stylized model, because rubric embeddings d-separate the black-box embeddings and the ground truth, conditional on the rubric embeddings, any function of the black-box embeddings—*a fortiori*, any predictions based on them—should be uncorrelated with the true labels. In fact, this is precisely what we observe: the center and right-hand panels of Fig. 3 illustrate that while kitchen sink and black-box models achieve lower RMSE than the rubric embedding models relative to the *proxy* label, relative to the ground truth, rubric embedding models achieve the highest accuracy. (See Fig. A2 for the close agreement between theoretical and empirically predicted values of the MSE gap.)

4 Discussion

Label bias is a long-standing and pernicious problem in designing equitable algorithms. But despite its importance, there are few general techniques available to address it—short of collecting less biased labels [10]. Rubric embeddings provide a practical approach to addressing this challenge. Our analytical and empirical results demonstrate that rubric embeddings can yield more accurate and equitable decisions. Our results further suggest that bias arising from proxy labels is not only

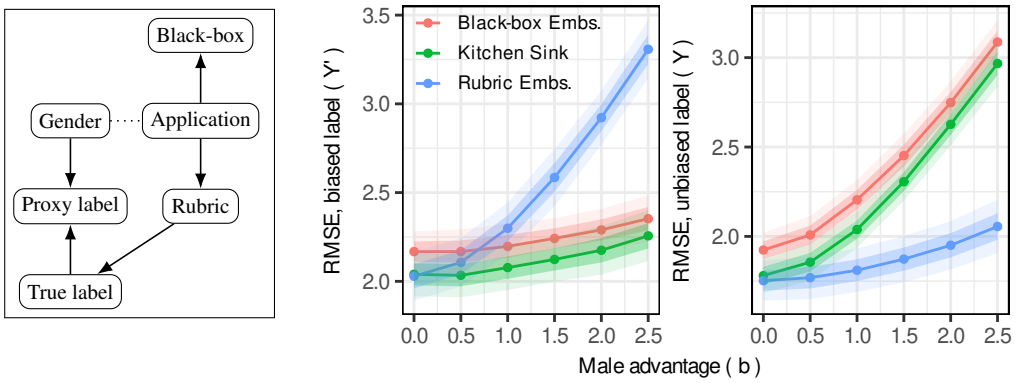

Figure 3: *Theoretical explanation and empirical verification of rubric embedding models' superior predictive performance.* Left: *A causal DAG representing our admissions setting. The dotted line indicates correlation.* Center: *RMSEs of rubric embedding, black-box embedding, and kitchen sink models, evaluated relative to proxy labels Y'.* Right: *RMSEs of rubric embedding, black-box embedding, and kitchen sink models, evaluated relative to ground truth Y.*

a property of the labels themselves, but also of the representations used to model them. There are, however, important limitations of our approach, three of which we discuss below.

First, constructing a rubric requires substantial time and domain expertise. The ideal rubric includes all legitimate decision-relevant factors and excludes irrelevant ones. In our admissions setting, we developed the rubric through iterative engagement with admissions officers, including reviewing evaluation guidelines, observing evaluation processes, and participating in admissions committee discussions. We then used LLMs to operationalize these criteria into a structured representation that could be consistently applied to applications, followed by manual review to resolve discrepancies between human and algorithmic assessments. While this process is difficult—and can only partially be automated—we believe it is critical for developing a high-quality rubric. An additional benefit is that it can prompt substantive discussion among decision-makers about which factors should be considered in the first place.

Second, basing predictions on rubric embeddings mitigates some forms of bias but not all. By construction, rubric embeddings enumerate factors that are considered legitimate for decision-making, and exclude group membership. As a result, they offer strong protection against *disparate treatment*, the form of discrimination typically studied in correspondence experiments. However, even when the rubric captures the appropriate set of decision-relevant factors, label bias can distort the weights assigned to those factors. For example, if female applicants are more likely to engage in public service, then labels biased against women may lead to an inappropriately low weight on public service, thereby penalizing qualified women. This form of *disparate impact* is more difficult to eliminate. In practice, rubric embeddings mitigate this risk by constraining the space of admissible models to those based on semantically meaningful factors. Consistent with this intuition, Fig. 2 shows that the gender composition of selected applicants remains stable even under substantial label bias.

Third, by design, rubric embeddings exclude group membership and related features from the representation. While this provides protection against disparate treatment, it also limits the ability of models to incorporate group-specific information, even in settings where such information may be relevant or normatively justified. For example, in some medical contexts, group membership can be predictive of risk due to underlying structural or biological differences, and incorporating such information may improve outcomes [48]. Similarly, in policy settings such as affirmative action, decision-makers may wish to explicitly account for group membership to address historical inequities. By restricting attention to group-neutral representations, rubric embeddings may preclude these types of interventions. This limitation reflects a fundamental tension between enforcing group-blind decision rules and allowing for context-dependent uses of group information.

Despite these limitations, rubric embeddings offer a practical approach for mitigating label bias in a variety of high-stakes settings, from university admissions to hiring. More broadly, our results suggest that addressing bias requires not only improving data and learning algorithms, but also

reconsidering how inputs are represented. By grounding predictions in semantically meaningful, domain-specific criteria, rubric embeddings provide a principled way to align learned models with the underlying construct of interest, offering a promising direction for designing more reliable and equitable decision-making systems.

References

[1] Jeffrey Dastin. Insight - Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, October 2018. URL https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/.

[2] Alex Chohlas-Wood, Madison Coots, Sharad Goel, and Julian Nyarko. Designing equitable algorithms. *Nature Computational Science*, 3, 2023.

[3] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29:3315–3323, 2016.

[4] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012.

[5] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of Innovations in Theoretical Computer Science (ITCS)*, 2017.

[6] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pages 134–148, 2018.

[7] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

[8] Sam Corbett-Davies, Johann D. Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. The Measure and Mismeasure of Fairness. *Journal of Machine Learning Research*, 24(312): 1–117, 2023. ISSN 1533-7928. URL http://jmlr.org/papers/v24/22-1511.html.

[9] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806, 2017.

[10] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[11] Heinrich Jiang and Ofir Nachum. Identifying and Correcting Label Bias in Machine Learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pages 702–712. PMLR, June 2020. URL https://proceedings.mlr.press/v108/jiang20a.html.

[12] Jialu Wang, Yang Liu, and Caleb Levy. Fair Classification with Group-Dependent Label Noise. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 526–536, New York, NY, USA, March 2021. Association for Computing Machinery. ISBN 978-1-4503-8309-7. doi: 10.1145/3442188.3445915. URL https://dl.acm.org/doi/10.1145/3442188.3445915.

[13] Michael Zanger-Tishler, Julian Nyarko, and Sharad Goel. Risk scores, label bias, and everything but the kitchen sink. *Science Advances*, 10(13):eadi8411, March 2024. doi: 10.1126/sciadv.adi8411. URL https://www.science.org/doi/10.1126/sciadv.adi8411.

[14] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept Bottleneck Models. In *Proceedings of the 37th International Conference on Machine Learning*, pages 5338–5348. PMLR, November 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

[15] Zhen Tan, Lu Cheng, Song Wang, Yuan Bo, Jundong Li, and Huan Liu. Interpreting Pretrained Language Models via Concept Bottlenecks (Extended Abstract). volume 12, pages 10942–10946, September 2025. doi: 10.24963/ijcai.2025/1221. URL https://www.ijcai.org/proceedings/2025/1221.

[16] Chung-En Sun, Tuomas Oikarinen, Berk Ustun, and Tsui-Wei Weng. Concept Bottleneck Large Language Models, September 2025. URL http://arxiv.org/abs/2412.07992. arXiv:2412.07992 [cs].

[17] Josh Magnus Ludan, Qing Lyu, Yue Yang, Liam Dugan, Mark Yatskar, and Chris Callison-Burch. Interpretable-by-Design Text Understanding with Iteratively Generated Concept Bottleneck, April 2024. URL http://arxiv.org/abs/2310.19660. arXiv:2310.19660 [cs].

[18] Ying-Chun Lin, Jennifer Neville, Jack Stokes, Longqi Yang, Tara Safavi, Mengting Wan, Scott Counts, Siddharth Suri, Reid Andersen, Xiaofeng Xu, Deepak Gupta, Sujay Kumar Jauhar, Xia Song, Georg Buscher, Saurabh Tiwary, Brent Hecht, and Jaime Teevan. Interpretable user satisfaction estimation for conversational systems with large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11100–11115, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.598. URL https://aclanthology.org/2024.acl-long.598/.

[19] Vojtěch Balek, Gustav Sourek, and Tomáš Kliegr. LLM-based feature generation from text for interpretable machine learning, 2024. URL https://arxiv.org/abs/2409.07132.

[20] Ilker Demirel, Lawrence Shi, Zeshan Hussain, and David Sontag. LLMs can construct powerful representations and streamline sample-efficient supervised learning, 2026. URL https://arxiv.org/abs/2603.11679.

[21] An Yan, Yu Wang, Yiwu Zhong, Zexue He, Petros Karypis, Zihan Wang, Chengyu Dong, Amilcare Gentili, Chun-Nan Hsu, Jingbo Shang, and Julian McAuley. Robust and Interpretable Medical Image Classifiers via Concept Bottleneck Models, October 2023. URL http://arxiv.org/abs/2310.03182. arXiv:2310.03182 [cs].

[22] Jihye Choi, Jayaram Raghuram, Yixuan Li, and Somesh Jha. Adaptive concept bottleneck for foundation models under distribution shifts, 2024. URL https://arxiv.org/abs/2412.14097.

[23] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

[25] Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*, 2022.

[26] Johann D. Gaebler, Sharad Goel, Aziz Huq, and Prasanna Tambe. Auditing large language models for race & gender disparities: Implications for artificial intelligence–based hiring. *Behavioral Science & Policy*, 2025. doi: 10.1177/23794607251320229.

[27] Marianne Bertrand and Esther Duflo. Field experiments on discrimination. *Handbook of economic field experiments*, 1:309–393, 2017.

[28] S Michael Gaddis. Understanding the "how" and "why" aspects of racial-ethnic discrimination: A multimethod approach to audit studies. *Sociology of Race and Ethnicity*, 5(4):443–455, 2019.

[29] Marianne Bertrand and Sendhil Mullainathan. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, 94(4):991–1013, 2004.

[30] Microsoft. Azure AI Document Intelligence, 2024. URL https://azure.microsoft.com/en-us/products/ai-services/ai-document-intelligence.

[31] OpenAI. OpenAI API. https://platform.openai.com/docs/guides/embeddings, 2023. Accessed on October 14, 2024.

[32] Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, et al. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35: 30233–30249, 2022.

[33] SFFA v. Harvard. Students for Fair Admissions, Inc., Petitioner, v. President and Fellows of Harvard College. Students for Fair Admissions, Inc., Petitioner, v. University of North Carolina, et al., 2023. https://www.supremecourt.gov/opinions/22pdf/20-1199_l6gn.pdf.

[34] J. Kenneth Tay, Balasubramanian Narasimhan, and Trevor Hastie. Elastic net regularization paths for all generalized linear models. *Journal of Statistical Software*, 106(1):1–31, 2023. doi: 10.18637/jss.v106.i01.

[35] Donald B Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York, 1987.

[36] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

[37] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7237–7256, 2020.

[38] Harrison Edwards and Amos Storkey. Censoring representations with an adversary. In *Proceedings of the International Conference in Learning Representations*, 2016.

[39] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[40] Coalition for TJ v. Fairfax County School Board. Coalition for TJ v. Fairfax County School Board, 2023. 68 F.4th 864 (4th Cir. 2023). https://law.justia.com/cases/federal/appellate-courts/ca4/22-1280/22-1280-2023-05-23.html.

[41] Boston Parent Coalition v. Boston School Committee. Boston Parent Coalition for Academic Excellence Corp. v. School Committee for the City of Boston, 2023. 89 F.4th 46 (1st Cir. 2023). https://law.justia.com/cases/federal/appellate-courts/ca1/21-1303/21-1303-2023-12-19.html.

[42] Jiaxin Wen, Zachary Ankner, Arushi Somani, Peter Hase, Samuel Marks, Jacob Goldman-Wetzler, Linda Petrini, Henry Sleight, Collin Burns, He He, Shi Feng, Ethan Perez, and Jan Leike. Unsupervised elicitation of language models, 2025. URL https://arxiv.org/abs/2506.10139.

[43] Prasanna Parasurama and João Sedoc. Gendered language in resumes and its implications for algorithmic bias in hiring. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, Seattle, Washington, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.gebnlp-1.7. URL https://aclanthology.org/2022.gebnlp-1.7/.

[44] Crystal Yang and Will Dobbie. Equal Protection Under Algorithms: A New Statistical and Legal Framework. *Michigan Law Review*, 119(2):291–396, November 2020. ISSN 0026-2234. doi: https://doi.org/10.36644/mlr.119.2.equal. URL https://repository.law.umich.edu/mlr/vol119/iss2/3.

[45] Devin G. Pope and Justin R. Sydnor. Implementing Anti-discrimination Policies in Statistical Profiling Models. *American Economic Journal: Economic Policy*, 3(3):206–231, August 2011. ISSN 1945-7731. doi: 10.1257/pol.3.3.206. URL https://www.aeaweb.org/articles?id=10.1257/pol.3.3.206.

[46] Yixin Wang, Dhanya Sridhar, and David M Blei. Equal opportunity and affirmative action via counterfactual predictions. *arXiv preprint arXiv:1905.10870*, 2019.

[47] OpenAI. GPT-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

[48] Madison Coots, Soroush Saghafian, David Kent, and Sharad Goel. Reevaluating the role of race and ethnicity in diabetes screening. *arXiv preprint arXiv:2306.10220*, 2023.

[49] OpenAI. GPT-5.4 Thinking system card. https://deploymentsafety.openai.com/gpt-5-4-thinking, 2026. Includes addendum on GPT-5.4 mini, March 17, 2026.

A Mathematical Appendix

Conventions and non-degeneracy assumptions. For an integrable random variable A and random object W, we write $\mathbb{E}[A \mid W = w]$ for a chosen measurable version of the conditional expectation as a function of w; by $\mathbb{E}[A \mid Z = \zeta(w)]$ we mean the composition of the maps $\mathbb{E}[A \mid Z = z]$ and $\zeta(w) : \mathcal{W} \to \mathcal{Z}$.

To avoid trivial degeneracies arising from the choice of version, we make the following simple regularity assumptions. First, the probability of belonging to either gender is positive, i.e.,

$$0 < \Pr(G = m) < 1. \tag{7}$$

Second, we assume that the distributions of $T(X, m)$ and $T(X, f)$ are absolutely continuous with respect to the distribution of X, i.e.,

$$\text{if} \quad \Pr(X \in E) = 0 \qquad \text{then} \quad \Pr(T(X, g) \in E) = 0, \quad g \in \{m, f\}. \tag{8}$$

All functions are assumed to be measurable and all random variables integrable. All equalities between random variables should be understood as almost sure equality.

Proofs. We begin with the proof of Prop. 2.

Proof of Prop. 2. By linearity and the tower property

$$\begin{aligned}\nu(X) &= \mathbb{E}[Y \mid r(X)] + \mathbb{E}[B \mid r(X)] \\ &= \mathbb{E}[Y \mid r(X)] + \sum_{g \in \{m,f\}} \mathbb{E}[B \mid r(X), G = g] \cdot \Pr(G = g \mid r(X)).\end{aligned}$$

Since $B \perp\!\!\!\perp X \mid G$, $B \perp\!\!\!\perp r(X) \mid G$, and so $\mathbb{E}[B \mid r(X), G = g] = \mathbb{E}[B \mid G = g]$. Define $\beta(g) \stackrel{\text{def}}{=} \mathbb{E}[B \mid G = g]$. Then

$$\nu(X) = \mu(X) + \beta(f) + [\beta(m) - \beta(f)] \cdot \pi(X),$$

where we have used the fact that $\Pr(G = f \mid r(X)) = 1 - \pi(X)$.

Substituting this expression into $\mathbb{E}[\nu(T(X, m)) - \nu(T(X, f))]$ yields

$$\mathbb{E}[\mu(T(X, m)) - \mu(T(X, f))] + [\beta(m) - \beta(f)] \cdot \mathbb{E}[\pi(T(X, m)) - \pi(T(X, f))],$$

which is Eq. (3). □

We next state and prove a straightforward but important consequence of orthogonalization.

Lemma A1. *If $R \perp\!\!\!\perp G$ and $\Pr(d(R) = 1) > 0$, then $\Pr(G = g \mid d(R) = 1) = \Pr(G = g)$.*

Proof. Since $R \perp\!\!\!\perp G$, it follows that any $\sigma(R)$-measurable function is independent of G, whence $d(R) \perp\!\!\!\perp G$, i.e., $\Pr(G = g \mid d(R) = 1) = \Pr(G = g)$ for $g \in \{m, f\}$. □

Both redaction and marginalization result in decision algorithms that are unbiased. In the case of marginalization, this depends on an intuitive consistency condition for the manipulation T. Note that manipulations $T(x, g)$ for different g typically mirror one another, altering names, pronouns, or other elements of x in parallel. Consequently, the result of applying a manipulation T multiple times is typically the same as the result of the final application:

$$T(T(X, g_0), g_1) \stackrel{d}{=} T(X, g_1), \qquad g_0, g_1 \in \{m, f\}. \tag{9}$$

Distributional equality accounts for stochasticity in the manipulation.

Proposition A2. *With the same notation as in Prop. 2, let $h(x) : \tilde{\mathcal{X}} \to \mathbb{R}$ be arbitrary.*

- ***Redaction**: If ρ and T satisfy Eq.* (5)*, then* $\mathrm{bias}_T(h \circ \rho) = 0$.
- ***Marginalization**: If T satisfies Eq.* (9)*, then* $\mathrm{bias}_T(h_T) = 0$.

Proof. First, we consider *redaction*. Observe that

$$\mathrm{bias}_T(h \circ \rho) = \mathbb{E}[h(\rho(T(X, m))) - h(\rho(T(X, f)))] = \mathbb{E}[h(\rho(X)) - h(\rho(X))] = 0$$

where the final equality follows from Eq. (5).

Next, we consider *marginalization*. Substituting Eq. (6) into Def. 1 gives that

$$\mathrm{bias}_T(h_T) = \mathbb{E}[h_T(T(X, m)) - h_T(T(X, f))].$$

Expanding using Eq. (6), this becomes

$$\mathbb{E}\big[w \cdot h(T(T(X, m), m)) + (1 - w) \cdot h(T(T(X, m), f)) \\ - w \cdot h(T(T(X, f), m)) - (1 - w) \cdot h(T(T(X, f), f))\big],$$

which, by the linearity of expectation, equals

$$w \cdot \mathbb{E}[h(T(T(X, m), m))] + (1 - w) \cdot \mathbb{E}[h(T(T(X, m), f))] \\ - w \cdot \mathbb{E}[h(T(T(X, f), m))] - (1 - w) \cdot \mathbb{E}[h(T(T(X, f), f))].$$

Recalling Eq. (9), this in turn equals

$$w \cdot \mathbb{E}[h(T(X, m))] + (1 - w) \cdot \mathbb{E}[h(T(X, f))] - w \cdot \mathbb{E}[h(T(X, m))] - (1 - w) \cdot \mathbb{E}[h(T(X, f))],$$

which is zero. □

B Audit study

To create the materials necessary to conduct our audit study, we iteratively process each type of application material (essays, letters of recommendation, and resumes) using an LLM-based pipeline. We use OpenAI's `gpt-5-mini` [47] for template construction and repair, and `gpt-5.4-mini` [49] for validation. (We do not create audit study materials for an applicant's transcripts, as the standardized text representations we create contain neither explicit references to gender nor applicants' names.)

Template construction. Our pipeline consists of three steps. First, we provide the language model with an application material's extracted markdown text, prompting it to return the otherwise unaltered full text with wrappers around all signals of gender and indications of an applicant's name in standardized templates formatted with ASCII control characters. For gender signals, these templates include placeholders with the original text, as well as the appropriate male, female, and gender-neutral alternatives. For example, the literal string `her` is transformed into `\x02her\x1Ftheir\x1Fhis\x1Ffher\x03`. For names, the model inserts a distinguished name placeholder, the field's original text, and the type of name indicator (e.g., email address, last name, social media handle). Under this name templating paradigm `@johndoe` becomes `@\x02NAME\x1Fhandle\x1Fjohndoe\x03`. See Prompt A1 for full details.

After a template has been generated, we use a different model to validate the correctness of the replacements. For the validation step, we prompt the model with the original prompt (Prompt A1), the template to be validated, and an additional prompt instructing it to identify and similarly replace any uncaptured signals of gender and applicant name. If no omissions or errors are found, the validator returns an empty list. See Prompt A2 for full details. In addition to LLM-driven validation checks, we confirm locally that the gender and name markers are properly formatted, and that the "original" variant is unchanged relative to the generating document using Python's `difflib`. Templates passing both LLM and local checks are considered complete.

Finally, we attempt to repair any errors identified during the validation step. We prompt the model with the generated template, a list of all the issues raised during the validation step, and instruct it to output a template with the listed errors resolved. The full instructions are given in Prompt A3. The output is then re-validated. When creating the materials for our audit study, we repeatedly validate and repair documents until a small number ($< .005\%$) of uncompleted files remained for manual review. Our manual review consisted of either manual approval of model false-positives or human resolution of outstanding issues.

Generating transformed materials. Once a valid template has been constructed, we instantiate four variants—male, female, neutral, and "original"—by filling in the appropriate gender and name placeholders. For gender, the original variant preserves the applicant's submitted language, and the male, female, and gender-neutral variants use corresponding LLM-determined substitute. To appropriately swap in a new name for an applicant for their male and female variants, we compiled the names and home countries of historical applicants. In the male and female variants of applicants' materials, they are assigned random names belonging to an applicant of the appropriate gender from the applicants' home countries. In the neutral material variants, we substitute "FIRSTNAME LASTNAME," as appropriate, for name placeholders in the template and in the original, their true name. (Original variants are identical to the generating document and are created primarily to validate the template generation process.)

Generating transformed features. For each of the four document variants we generate for an applicant, we embed all the materials using OpenAI's `text-embedding-3-large` [31], yielding a collection of black-box embeddings for each variant. Similarly, we run our structured rubric extraction pipeline on each generated variant, producing rubric embedding features for all applicants across all conditions considered in our analysis.

This process yields the transformations $T(x, g)$ for $g \in \{m, f\}$ required for our audit study, as well as the gender-neutral materials required to evaluate the redaction technique described in Section 3.1.

Audit Study Templating Prompt

```
You are assisting in a correspondence (audit) experiment on admissions
to a public policy master's program. Reproduce the input document
character-for-character, with each gendered signal and each piece of
the applicant's identifying information wrapped in an inline marker
so downstream code can render gender-swapped variants.

OUTPUT: only the templated document. No preamble, no commentary, no
JSON wrapper, no code fences. Begin your output with the very first
character of the input document (or with a marker, if the document
begins with a name).

MARKER FORMAT

Markers are delimited by ASCII control characters:

  STX = \x02 (Start of Text)
  ETX = \x03 (End of Text)
  US  = \x1F (Unit Separator)

Gender marker - exactly four fields, separated by US:

  \x02original\x1Fneutral\x1Fmale\x1Ffemale\x03

The four fields are: the source span verbatim, the gender-neutral
variant, the male variant, and the female variant. If a particular
variant truly has no reasonable analogue (e.g. "her pregnancy" has
no clean male variant), use the literal token  null  in that field
(no quotes). Use null sparingly - almost everything has an analogue.

Name marker - exactly three fields, the first being the literal
string "NAME":

  \x02NAME\x1Fpart\x1Foriginal\x03

`part` MUST be EXACTLY one of:
  first  last  nickname  other  email  handle  url

The third field is the exact name token from the source.

INTEGRITY RULE - IMPORTANT

If you remove every marker and replace it with its `original` field
(field 1 of gender markers; field 3 of name markers), the result MUST
equal the input document byte-for-byte. Do NOT fix typos. Do NOT
normalize whitespace. Do NOT change anything outside a marker. The
input may contain markdown, HTML comments, page-break artifacts -
preserve all of it.

WHAT TO WRAP IN A GENDER MARKER

Every occurrence of an explicit gender signal that refers to the
APPLICANT (not third parties):
  - Pronouns: he/she, him/her, his/her, hers, himself/herself.
    Use "themself"/"themselves" for the neutral reflexive.
  - Gendered titles: Mr., Ms., Mrs., Miss
  - Gendered nouns: actress, chairman, businesswoman
  - Family roles applied to the applicant: father, daughter
  - Explicit gender statements: "as a woman", "being male"
  - Gendered modifiers: women's varsity team
  - Membership in gender-specific groups (sororities/fraternities,
    Boy/Girl Scouts, single-sex schools). Use a reasonable
    opposite-gender analogue (Eagle Scout ↔ Girl Scout Gold Award;
    Alpha Phi sorority ↔ a plausible fraternity).
```

```
WHAT NOT TO WRAP IN A GENDER MARKER

  - Names, emails, URLs, handles - wrap with NAME markers instead
  - Profession/interest correlates: nursing, engineering, ballet
  - Non-gendered honorifics: Dr., Prof., Rev.
  - Non-gendered pronouns: my, I, our
  - Gender signals referring to a third party (recommender, family,
    colleague). Leave them entirely outside any marker.

WHAT TO WRAP IN A NAME MARKER

Every occurrence of the APPLICANT'S name or an identifier that leaks
the name. Wrap each occurrence individually. When a third party
shares a name token (e.g. recommender shares the applicant's first
name), wrap ONLY the applicant's occurrences and leave the third
party's bare.

EXAMPLE

Input:
  "Jane Smith excelled. She wrote her thesis on social welfare."

Output:
  "\x02NAME\x1Ffirst\x1FJane\x03 \x02NAME\x1Flast\x1FSmith\x03 excelled.
  ↪  \x02She\x1FThey\x1FHe\x1FShe\x03 wrote \x02her\x1Ftheir\x1Fhis\x1Fher\x03
  ↪  thesis on social welfare."
```

Prompt A1: *The templating prompt, which instructs the model to wrap all applicant gender signals and name tokens in structured inline markers.*

Audit Study Validation Prompt

```
YOU ARE NOW REVIEWING the templated output that follows. The original
model produced this output by applying the rules above to a source
document. Your job is to find any APPLICANT gender signal that was
left outside a marker - i.e., anywhere the original model failed to
follow rule 2 above.

OUTPUT a JSON object with a single field:


{"missed": ["...", "...", ...]}

Each entry should be a short string identifying the missed signal and
its surrounding context, e.g. "her in 'complete her thesis'".

Return an EMPTY list if nothing was missed.

DO NOT flag:

  - Gender signals referring to third parties (recommender, family
    members, named non-applicants, colleagues, etc.). These are
    correctly left outside markers.
  - Non-gendered honorifics (Dr., Prof., Rev.)
  - Profession/interest correlates (nursing, ballet, engineering, etc.)
  - Generic references that are not tied to the applicant

Be strict but not paranoid: only flag tokens you are confident the
applicant of this document possesses or is described by.
```

Prompt A2: *The validation prompt, which instructs the model to confirm if the prior templating step missed any indicators of gender or applicant name.*

Audit Study Repair Prompt

```
This template was flagged as potentially still containing incomplete
processing. The issues identified:

{issues_block}

Some of these flags may be false positives - use your judgment:

- If a flagged token correctly refers to a third party (recommender,
  family member, colleague, etc.), leave the existing template
  unchanged for that token.
- If a flagged token genuinely refers to the applicant and should
  have been wrapped, add the appropriate marker.
- For parse errors, fix the malformed marker(s) following the format
  rules in the system prompt.
- For integrity errors, restore the missing/modified text outside any
  marker so the marker-stripped reconstruction matches the source.

Output the full corrected template, no commentary.

PREVIOUS TEMPLATE:

{template}
```

Prompt A3: *The repair prompt, which instructs the model to repair any of the issues raised in the validation step.*

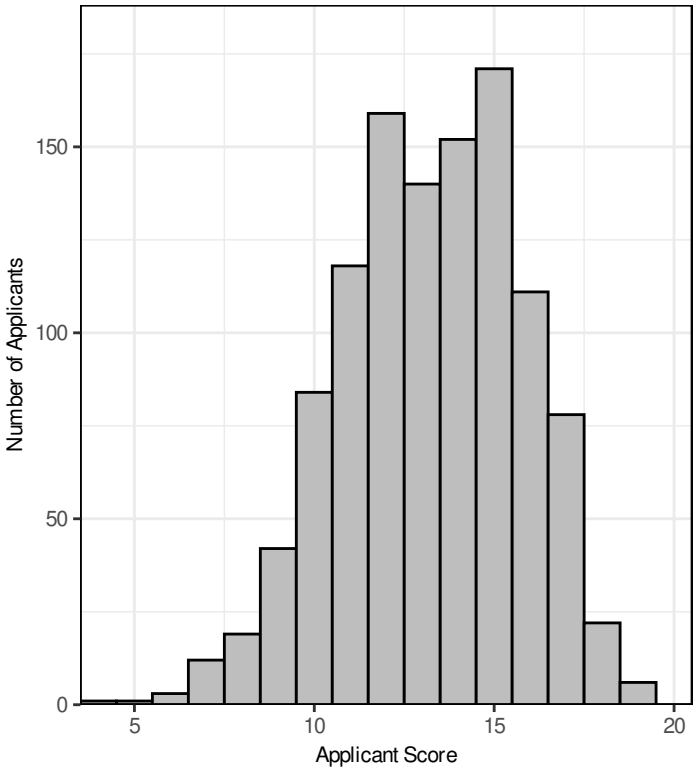

Figure A1: *The distribution of actual applicant scores in the 2025–2026 round of admissions.*

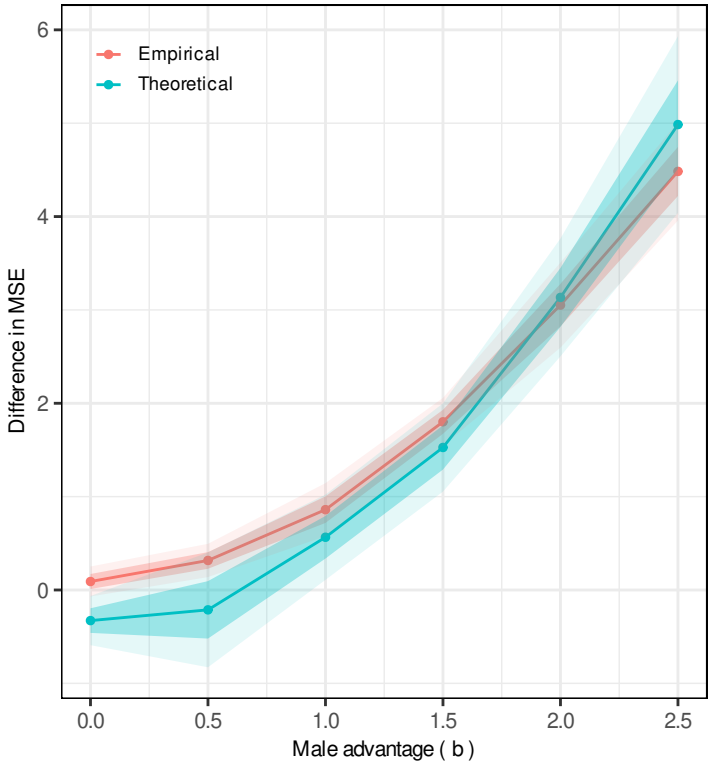

Figure A2: *Theoretical predicted and empirical values of the gap in MSE between a kitchen sink and rubric embeddings model.*

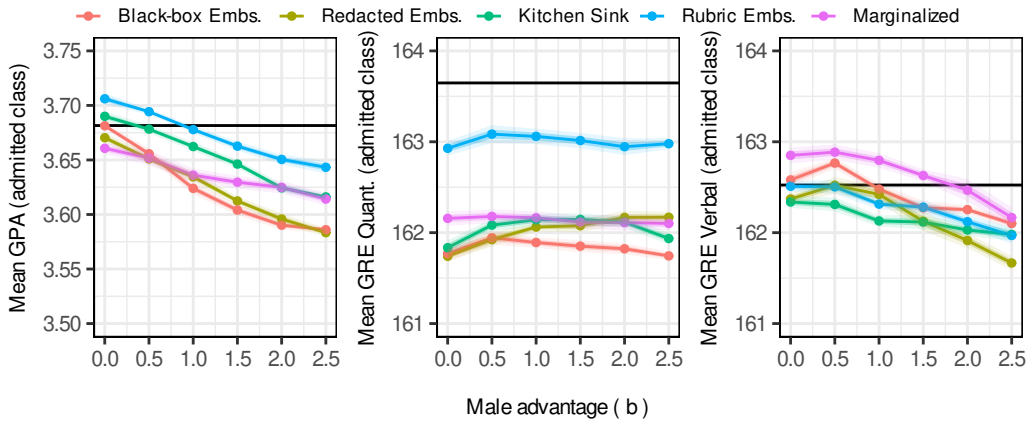

Figure A3: *The impact on various measures of the strength of the admitted class using models of the proxy label with the corresponding level b. Colors indicate the bias mitigation technique. Dark and light shaded regions show pointwise 68% and 95% confidence intervals, respectively. Solid lines show corresponding values for the top 20% of students admitted according to actual scores Y. Left panel: The average GPA (graduate and undergraduate) of the admitted class. Center panel: The average GRE quantitative score of the admitted class. Right panel: The average GRE verbal score of the admitted class.*