Mitigating Label Bias with Interpretable Rubric Embeddings

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

Johann D. Gaebler
Stern School of Business
New York University
New York, NY 10012
jgaeb@stern.nyu.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. 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 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.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

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 practical strategy for mitigating label bias in a wide range of decision-making contexts, addressing a long-standing challenge in the equitable design of algorithms [1]. 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.

Analysis code and model outputs needed to reproduce the reported figures and statistics are available at `https://github.com/calisley/label-bias`.

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., 1–8]. 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 [9]. 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 [10], or adjusting the loss to account for group-dependent label noise [11]. 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. [12] 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 [13]. 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 [14–16]. 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 [17–19]. 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. [16] and Demirel et al. [19] both use LLMs to generate a collection of interpretable, semantically relevant features. Further, human-graded rubrics have been used to establish ground-truth resume quality in audit studies of LLM evaluators [20]. This literature has focused primarily on interpretability, sample efficiency, and robustness to distribution shift [21, 22]. However, to our knowledge, the use of concept-based representations or rubrics 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.

Specifically, we 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.

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

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

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

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*

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

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,

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

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.

Eq. (3) further indicates that $\text{bias}_T(\nu)$ depends on the baseline bias, $\text{bias}_T(\mu)$, of the algorithm when predicting the unbiased, ground-truth label Y. It is perhaps surprising that an algorithm can exhibit such bias even when trained on an unbiased label. Prop. 3 below sheds light on this phenomenon by deriving an alternative expression for the bias of an algorithm. For clarity and tractability, we state the result for linear predictors but the intuition holds more broadly.

Proposition 3. *Let $X \in \mathcal{X}$ and $R = r(X) \in \mathbb{R}^p$ be as above, and let $Y \in \mathbb{R}$ be square-integrable. Let $h_R(r) = \alpha_0 + r^\top \alpha$ be the population OLS predictor of Y from R. Define the audit-induced representation shift*

$$s \stackrel{\text{def}}{=} \mathbb{E}[r(T(X, m)) - r(T(X, f))]\,.$$

Then

$$\text{bias}_T(h_R) = s^\top \text{Var}(R)^{-1} \text{Cov}(R, Y). \quad (5)$$

Prop. 3 illustrates that when features shift under the gender manipulation T, bias arises to the extent that those features are predictive of Y. Writing the population coefficient vector as $\alpha = \text{Var}(R)^{-1} \text{Cov}(R, Y)$, Eq. (5) becomes $\text{bias}_T(h_R) = s^\top \alpha$, meaning that bias is nonzero precisely when the average representation shift s has a component along the predictive direction α. As a result, omitted variable bias can lead to audit study bias, even in the presence of unbiased labels. As a simple illustration, note that when R is whitened (i.e., when $\text{Var}(R) = I_p$), Eq. (5) further simplifies to $s^\top \text{Cov}(R, Y)$, meaning that each feature individually contributes to the bias to the extent that it correlates with the outcome and changes under the transformation. If some feature relevant to Y is correlated with gender (for instance, if female applicants are more likely to engage in public service) but is not captured by R, then the components of R that *are* captured and that shift with gender act as proxies for the omitted feature, inducing bias.

The distinction between omitted variable bias and label bias is subtle, and they often lead to indistinguishable issues in practice. Consider the widely reported case of an internal hiring algorithm at

Amazon that penalized resumes containing terms like "women's" (e.g., as in "women's chess club captain") [30]. Conventional wisdom suggests that the penalty reflected biases in the historical hiring data. However, Proposition 3 suggests omitted variable bias as an alternative culprit. A model trained using data representations that omit job-relevant features correlated with gender would likely exhibit similar biases—a point we return to below.

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 [31]. We then separately embed each of the four document types (resumes, essays, letters, and transcripts) using OpenAI's `text-embedding-3-large` model [32]. Given the limited number of applications, we use only the first 250 dimensions of the embeddings for each category [33]. We additionally include some structured data, but we omit gender and other explicit demographic features because deployed admissions models likely cannot legally include them [34]. 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 [35]. 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 [36]. 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." 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 C.

By Eq. (4), $\mathrm{bias}_T(h)$ 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 0.91 ± 0.01, 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 $99.6\% \pm 0.3\%$, 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 $\mathrm{bias}_T(\mu)$ equals -0.64 ± 0.56, meaning that the algorithm is biased *against* men when predicting the ground-truth label. Proposition 3 provides one potential explanation: black-box embeddings fail to capture certain features relevant to admissions scores that are also correlated with gender. Gender subsequently serves as a proxy for that missing information, inducing the bias we observe with unbiased labels. We note that this quantity represents the bias of the *algorithm* and not the bias of the admissions *decisions*. Consistent with this interpretation, predictive models trained using a broader collection of features, including information about past student performance not well-captured by black-box embeddings, exhibit virtually no bias, 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 $\mathrm{bias}_T(h)$ (dashed line) broadly track the increase in empirical audit bias as the male advantage grows. [2] The middle and right panels illustrate the consequences for admissions. We consider top-k admissions policies that rank applicants

[2] The agreement is especially close because ridge regression models vary linearly in their training labels, so an analogue of Prop. 2 holds almost exactly even in finite samples; see Prop. A1 in Appendix A. For regression models that are not exactly linear in their labels, such as random forests, the agreement can be looser; see Fig. A5, middle row.

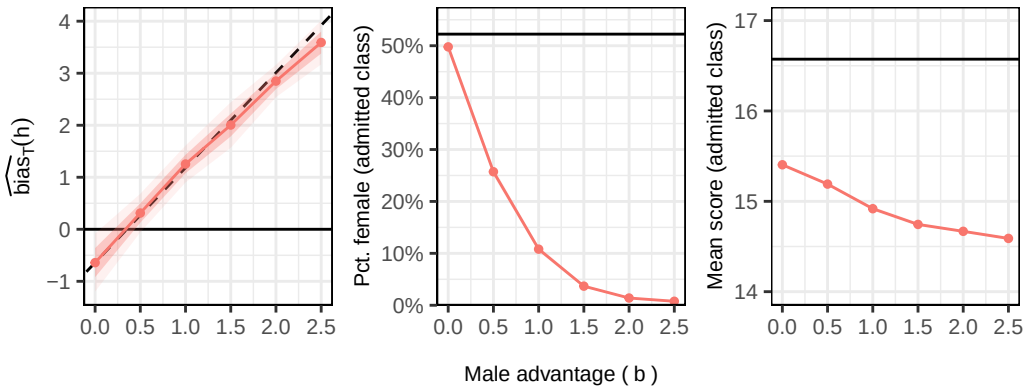

Figure 1: *Bias of regression models trained on proxy labels Y' and consequences for admitted classes. The x-axis in all panels denotes male advantage b in Eq.* (1)*. Solid black lines show zero bias (left) or corresponding values for the top 20% of students ranked by actual score Y (center, right).* 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. Dark and light shading show ± 1 and ± 2 standard errors, respectively.*

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 the propagation of label bias: black-box embeddings encode protected attributes such as gender, which in turn allows models trained on biased labels to reproduce those biases. Our theoretical analysis suggests this phenomenon stems in part from the ability of black-box embeddings to capture a rich set of signals in the data, including those related to gender. This suggests choosing representations that retain decision-relevant information while limiting gender signals. 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 [37, 38] aims to remove group information from a representation $R = r(X)$, ideally achieving $R \perp\!\!\!\perp G$. Orthogonalization and the closely related technique of adversarial debiasing [39, 40] 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 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" [34]; though cf. [41, 42].

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.,

$$\rho(T(X, g)) = \rho(X), \qquad g \in \{m, f\}. \tag{6}$$

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 C for full details.

Redaction is guaranteed to be unbiased with respect to T (see Prop. A3). 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 $81.3\% \pm 2.6\%$. 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, 43, 44]. 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., 45, 46], as well as the causal fairness literature [e.g., 47]. 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

$$h_T(x) \stackrel{\text{def}}{=} w \cdot h(T(x, m)) + [1 - w] \cdot h(T(x, f)), \tag{7}$$

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

Like redaction, this approach results in predictive algorithms with no bias with respect to T (see Prop. A3). 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, in collaboration with the admissions staff at our partner master's program, we develop a collection of 396 interpretable measures admissions officers deem substantively important in application review. See Appendix B for detailed information regarding the rubric construction process and final feature set.

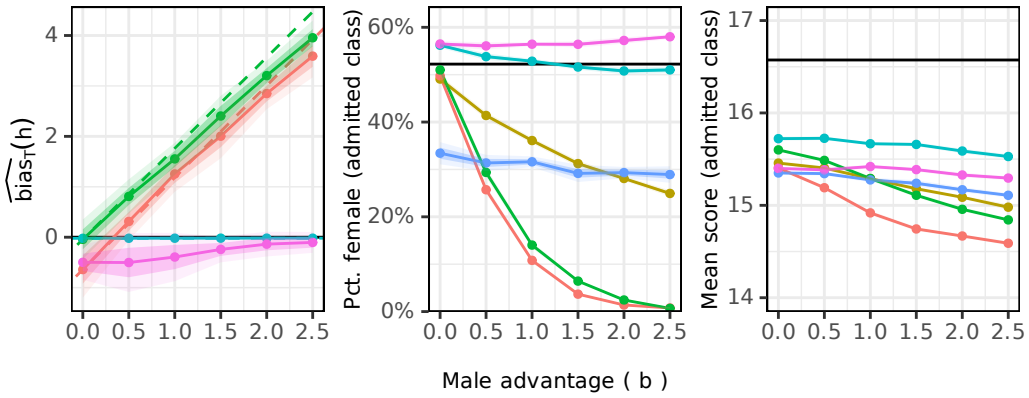

Figure 2: *Bias of models trained on proxy labels Y' under different bias mitigation techniques and consequences for admitted classes. The x-axis in all panels denotes male advantage b in Eq.* (1)*. Colors indicate the bias mitigation technique. Dark and light shading show ± 1 and ± 2 standard errors, respectively. 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.*

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 [48] to score the full applicant pool along each of the 396 dimensions.[3]

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 all 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"), (4) marginalized black-box embedding models ("marginalized"), (5) black-box embedding models in which each feature is uncorrelated with gender ("orthogonalized"). Although orthogonalization may itself raise legal concerns in jurisdictions that prohibit disparate treatment, we include it for completeness.

The audit bias and selected class composition results are shown in Fig. 2. As the left panel demonstrates, across levels of male advantage, we find virtually zero bias in rubric embedding models, consistent with their weak representation of gender (rubric-embedding models poorly predict gender, with an AUC of $61.7\% \pm 3.6\%$). We omit marginalized and redacted embedding models, which are guaranteed to have negligible bias. Consistent with our discussion in Section 3.1, we find orthogonalized models admit a lower-scoring cohort than rubric embedding models and exhibit moderate bias. This is likely due to the fact that orthogonalization removes gender information from the general-purpose embeddings and pushes decisions toward demographic parity, disregarding decision-relevant compositional differences between male and female applicants in our pool. 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. We also evaluate selected cohorts using prior GPA and GRE scores (Fig. A4). Rubric embeddings perform competitively across each of these measures, though they do not uniformly outperform the alternatives. Our results are robust to a variety of alternative specifications, including alternative model families, embedding models, different ground-truth labels and dimensions of advantage, and subsets of the rubric features. See Appendix D for further details.

[3]Fig. A2 reports average paired differences in re-extracted essay, resume, and recommendation-letter scores under the gender manipulation. Transcript-derived features and structured covariates are held fixed.

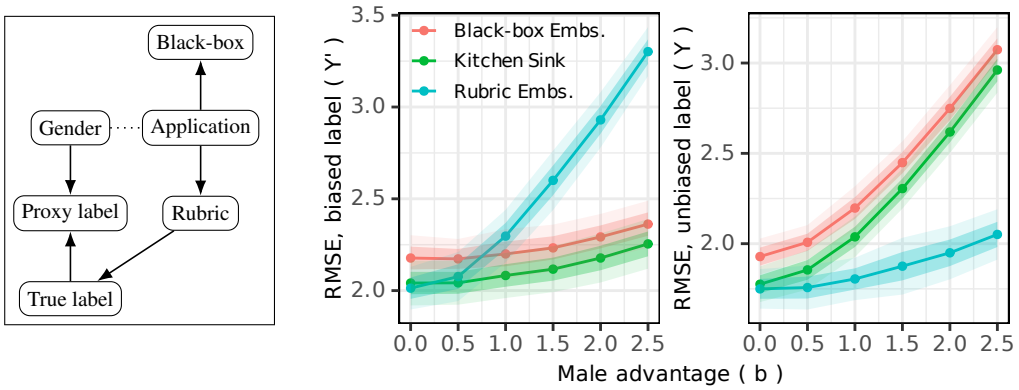

Figure 3: *Theoretical explanation and empirical verification of rubric embedding models' superior predictive performance.* Left: *A causal DAG representing our admissions setting. The dotted line indicates correlation.* Center: *RMSEs of rubric embedding, black-box embedding, and kitchen sink models, evaluated relative to proxy labels Y'.* Right: *RMSEs of rubric embedding, black-box embedding, and kitchen sink models, evaluated relative to ground truth Y. At $b = 0$, proxy labels retain zero-mean, unit-variance noise, so RMSEs against Y' and Y need not coincide. Dark and light shading show ± 1 and ± 2 standard errors, respectively.*

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 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, after adjusting for our rubric items, we find no significant gender disparity in admissions scores, with a male-female score difference of 0.17 (95% CI [-0.07, 0.41]). This suggests 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. [12] 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. 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. (Fig. A3 shows this 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 [9]. 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 a property of the labels themselves, but also of the representations used to model them. There are, however, important limitations of our approach, 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. In the future, more capable models may be able to increasingly facilitate or automate certain aspects of rubric design, though this may come at the expense of 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, rubric embeddings cannot erase genuine group differences in decision-relevant qualifications, and thus even when the rubric only contains decision-relevant information, label bias can still 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. In Appendix D, we explore this possibility by simulating label bias rewarding undergraduate economics, business, and engineering majors—a majority of whom are men—and find rubric embedding-based models show no audit-study bias and admit more gender-balanced classes than black-box embedding-based models. However, the quality of the admitted class declines as the advantage grows, since the rewarded features are directly represented in the rubric. More fundamentally, disparate impact is difficult to assess in this context because there is no consensus on which factors admissions should reward. For example, whether to weight leadership experience heavily is a context-dependent judgment.

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 [49]. 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.

Finally, our empirical analysis is limited to the singular empirical setting of admissions to a master's program. However, our population decomposition in Prop. 2 is domain agnostic and does not impose a parametric form on the conditional expectation. We therefore expect similar patterns to emerge wherever experts can clearly articulate evaluation criteria (e.g., hiring and language model alignment [50]) though confirming this empirically is important future work.

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.

Acknowledgments and Disclosure of Funding

This work received no external funding. The authors declare no competing interests.

References

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

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

[3] 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.

[4] 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.

[5] 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.

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

[7] 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`.

[8] 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.

[9] 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.

[10] 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`.

[11] 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`.

[12] 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`.

[13] 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`.

[14] 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`.

[15] 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].

[16] 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].

[17] 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/.

[18] 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.

[19] 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.

[20] Jiannan Xu, Gujie Li, and Jane Yi Jiang. AI self-preferencing in algorithmic hiring: Empirical evidence and insights, 2025. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5417394. SSRN 5417394.

[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] 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/.

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

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

[33] 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.

[34] 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.

[35] 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.

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

[37] 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.

[38] 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.

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

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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/.

[45] 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.

[46] 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.

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

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

[49] 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.

[50] Johann D Gaebler, Calvin Isley, Max Lamparth, Stephen Casper, and Sharad Goel. From constitutions to control: Interpretable rewards for aligning language models. *arXiv preprint arXiv:2609.33086*, 2026.

[51] Calvin Isley, Johann D. Gaebler, and Sharad Goel. Ai-written admissions essays are widespread but penalized, 2026. URL https://arxiv.org/abs/2609.22549.

[52] Johann D Gaebler, Calvin Isley, Christopher Avery, and Sharad Goel. Reassessing the role of standardized tests in university admissions. Working Paper 35407, National Bureau of Economic Research, July 2026. URL http://www.nber.org/papers/w35407.

[53] 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.

[54] Marvin N. Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17, 2017. doi: 10.18637/jss.v077.i01.

[55] Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. Open source strikes bread - new fluffy embeddings model, 2024. URL https://www.mixedbread.ai/blog/mxbai-embed-large-v1.

[56] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.emnlp-demos.6/.

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.,

$$0 < \Pr(G = m) < 1. \tag{8}$$

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.,

$$\text{if} \quad \Pr(X \in E) = 0 \qquad \text{then} \quad \Pr(T(X, g) \in E) = 0, \quad g \in \{m, f\}. \tag{9}$$

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). □

Proposition 2 characterizes the behavior of a predictive algorithm in the limit. In principle, in finite samples, predictive algorithms can deviate from Eq. (3). However, for ridge regression, the empirical analogue of Proposition 2 holds almost exactly even in finite samples, as we observe in Figures 1 and 2.

Proposition A1. *Let $\mathbf{R} \in \mathbb{R}^{n \times p}$ be the matrix with rows $r(X_1), \ldots, r(X_n)$, centered so that $\mathbf{R}^\top \mathbf{1} = 0$, and let $\mathbf{Y}, \mathbf{B} \in \mathbb{R}^n$ and $\mathbf{G} \in \{m, f\}^n$. Let $\mathbf{R}_m, \mathbf{R}_f \in \mathbb{R}^{N \times p}$ be the matrices with rows $r(T(\tilde{x}_j, m))$ and $r(T(\tilde{x}_j, f))$ for $\tilde{x}_1, \ldots, \tilde{x}_N \in \mathcal{X}$. For $\lambda > 0$, define for $\mathbf{h} \in \mathbb{R}^p$,*

$$\hat{h}(\mathbf{y}) \stackrel{\text{def}}{=} (\mathbf{R}^\top \mathbf{R} + \lambda I)^{-1} \mathbf{R}^\top \mathbf{y}, \qquad \widehat{\text{bias}}_T(\mathbf{h}) \stackrel{\text{def}}{=} \frac{1}{N} \left[\mathbf{1}^\top (\mathbf{R}_m - \mathbf{R}_f) \mathbf{h} \right].$$

Let $\hat{\beta}(g)$ be the average of B_i over i such that $G_i = g$, and suppose $\mathbf{R}^\top \boldsymbol{\varepsilon} = 0$, where $\varepsilon_i \stackrel{\text{def}}{=} B_i - \hat{\beta}(G_i)$. Then

$$\widehat{\text{bias}}_T(\hat{h}(\mathbf{Y} + \mathbf{B})) = \widehat{\text{bias}}_T(\hat{h}(\mathbf{Y})) + [\hat{\beta}(m) - \hat{\beta}(f)] \cdot \frac{1}{N} \sum_{j=1}^{N} [\hat{\pi}(T(\tilde{x}_j, m)) - \hat{\pi}(T(\tilde{x}_j, f))],$$

where $\hat{\pi}(x) \stackrel{\text{def}}{=} r(x)^\top \hat{h}(\mathbf{1}_m)$ and $\mathbf{1}_m$ is the indicator vector of $G_i = m$.

Proposition A1 states that for the ridge regression $\hat{h}(\mathbf{y})$ fit with design matrix $\mathbf{R}$ to labels $\mathbf{y}$, the empirical audit bias on a (possibly distinct) collection of applications $\tilde{x}_1, \ldots, \tilde{x}_N$ decomposes exactly as in Eq. (3), with π replaced by a linear ridge probability model predicting gender and fit to the same features and the independence condition replaced by orthogonality. The proof is almost identical to Proposition 2.

Proof of Prop. A1. By definition, $\mathbf{B} = \hat{\beta}(f)\mathbf{1} + [\hat{\beta}(m) - \hat{\beta}(f)] \cdot \mathbf{1}_m + \boldsymbol{\varepsilon}$. Since $\hat{h}(\mathbf{y})$ is linear in $\mathbf{y}$, $\hat{h}(\mathbf{1}) = 0$ because $\mathbf{R}$ is centered, and $\hat{h}(\boldsymbol{\varepsilon}) = 0$ because $\mathbf{R}^\top \boldsymbol{\varepsilon} = 0$. Therefore

$$\hat{h}(\mathbf{Y} + \mathbf{B}) = \hat{h}(\mathbf{Y}) + [\hat{\beta}(m) - \hat{\beta}(f)] \cdot \hat{h}(\mathbf{1}_m).$$

Since $\widehat{\mathrm{bias}}_T(\mathbf{h})$ is linear in $\mathbf{h}$, and $\widehat{\mathrm{bias}}_T(\hat{h}(\mathbf{1}_m)) = \frac{1}{N}\sum_j [\hat{\pi}(T(\tilde{x}_j, m)) - \hat{\pi}(T(\tilde{x}_j, f))]$ by the definition of $\hat{\pi}$, this gives the result. □

Proposition 3 follows as a straightforward consequence of the linearity of expectation.

Proof. By the definition of h_R, we have that $\mathrm{bias}_T(h_R)$ is

$$\mathbb{E}[\alpha_0 + r(T(X, m))^\top \alpha - \alpha_0 - r(T(X, f))^\top \alpha] = \mathbb{E}[r(T(X, m)) - r(T(X, f))]^\top \alpha,$$

where the equality follows from the linearity of expectation. The latter expression is $s^\top \alpha$. Since $\alpha = \mathrm{Var}(R)^{-1} \mathrm{Cov}(R, Y)$, it follows that

$$\mathrm{bias}_T(h_R) = s^\top \mathrm{Var}(R)^{-1} \mathrm{Cov}(R, Y),$$

which is Eq. (5). □

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

Lemma A2. *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:

$$T(T(X, g_0), g_1) \overset{d}{=} T(X, g_1), \qquad g_0, g_1 \in \{m, f\}. \tag{10}$$

Distributional equality accounts for stochasticity in the manipulation.

Proposition A3. *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.* (6)*, then* $\mathrm{bias}_T(h \circ \rho) = 0$.
- ***Marginalization**: If T satisfies Eq.* (10)*, 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. (6).

Next, we consider *marginalization*. Substituting Eq. (7) 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. (7), 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. (10), 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. □

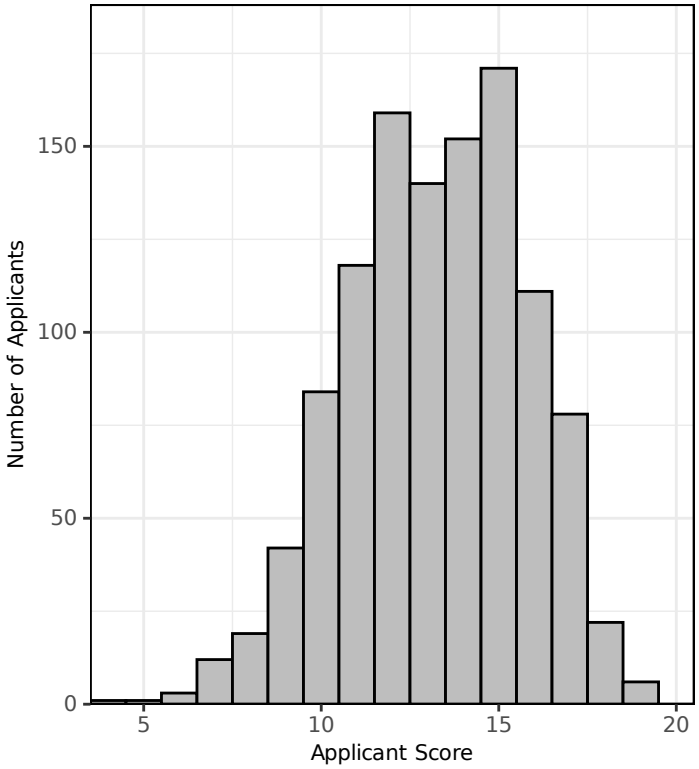

Figure A1: *The distribution of actual applicant scores in the 2025–2026 round of admissions.*

B Creating the rubric embeddings

While making admissions decisions, the officers at the school we study use a scoring guide to calibrate where they place applicants on the 4–20 scale. Fig. A1 shows the distribution of these scores. The scoring guide decomposes the final score into several subcategories, and describes in detail what separates applicants receiving each point value within each category, with reference to specific application features. Our original rubrics were based off of this scoring document: for each type of unstructured data applicants provide (e.g., transcripts) we identify where in the scoring guide that document is mentioned, and what information from that document is considered (e.g., number of prior quantitative courses).

Having enumerated each of these decision-relevant variables (or, the information necessary to construct them post-hoc) we then, for each type of material (resume, transcript, letter of recommendation, essay) prompt `gpt-5-mini` [48] to extract this information from the applicant's provided material(s), and return the result in a structured format with `gpt-5-mini`'s structured output functionality. We provide `gpt-5-mini` with a markdown version of the applicant's material, submitted as a PDF, converted using Microsoft Document Intelligence [31], along with the rubric and the expected structured output format. The resulting output is then converted to a structured format, aggregated to the applicant level, and included in our rubric embedding models. For further information regarding our document processing pipeline, see the appendices of concurrent work [51, 52].

Crucially, this process involved multiple rounds of iteration with admissions officers, observing their decision making process, and exploring what decision-relevant information we had missed when our preliminary model scores dramatically differed from their judgment. This iterative process grants us the 396 features we mention above. Of these 396 features, 14 are extracted from the applicant's essays, covering things such as grammatical correctness, clarity of writing, prompt adherence and substantive appropriateness. Eight measures come from resumes, summarizing applicants' work histories, including tenure in different sectors along with measures of seniority and career trajectory. We extract 39 measures from an applicant's letters of recommendation, including the strength of the recommender's endorsement (both overall and along a variety of personal dimensions),

references to specific program-relevant technical skills, and substantive appropriateness of the choice of recommender. Finally, the remaining 335 distinct measures are extracted from transcripts, covering 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). Applicants were required to submit five essays and three letters of recommendation, and many submitted both undergraduate and graduate transcripts. For essays and letters of recommendation, we take averages across each document-level extraction to aggregate to the applicant level. For transcripts, we encode each transcript dimension for undergraduate and graduate school independently (but average within each level of school.)

C 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` [48] for template construction and repair, and `gpt-5.4-mini` [53] 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\x1Fher\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 the 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.)

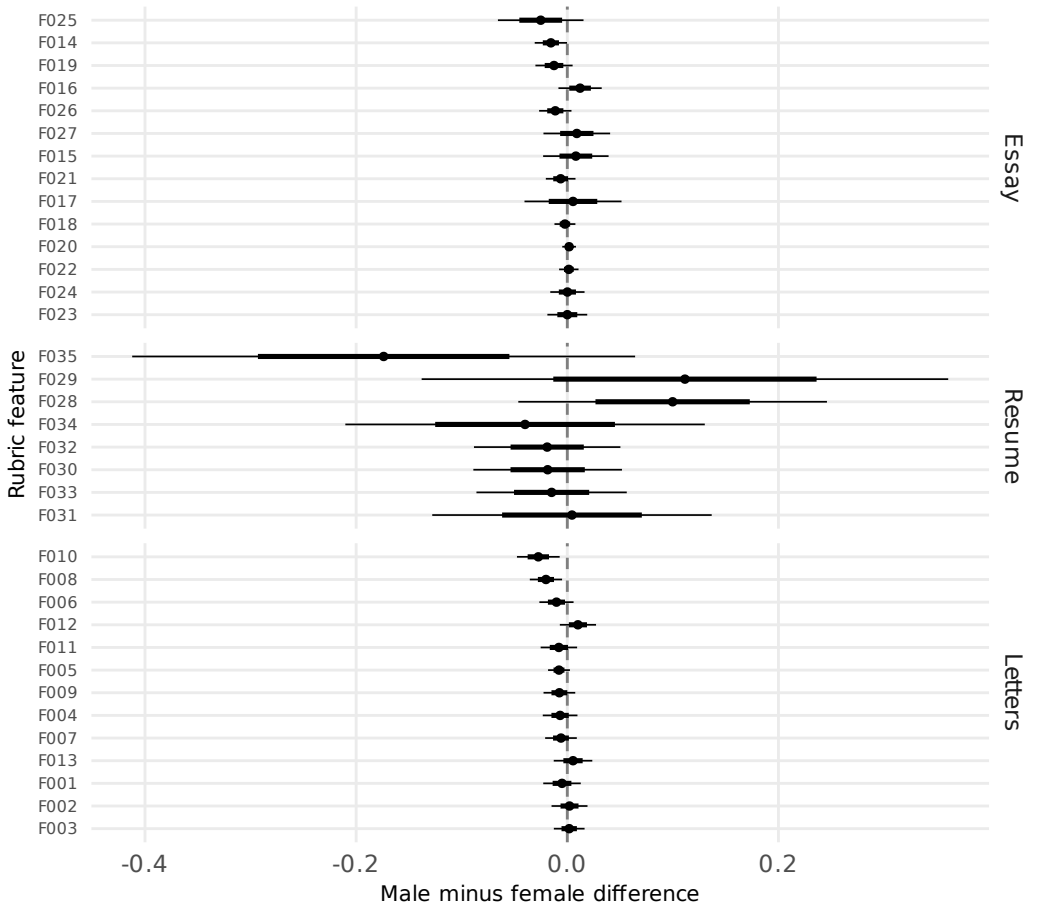

Figure A2: *Mean paired differences between male- and female-presenting applications in re-extracted essay, resume, and recommendation-letter features. Points show mean differences; thick and thin bars show* ± 1 *and* ± 2 *standard errors, respectively. Feature identifiers are anonymized. Transcripts and structured covariates are held fixed and omitted.*

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` [32], 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.

Our approach assumes the use of an LLM extractor that scores the rubric criteria consistently and accurately. Fig. A2 reports average paired differences in the re-extracted rubric features. Most estimated differences are close to zero, although some features exhibit detectable shifts. We do not replicate our rubric extraction with another LLM.

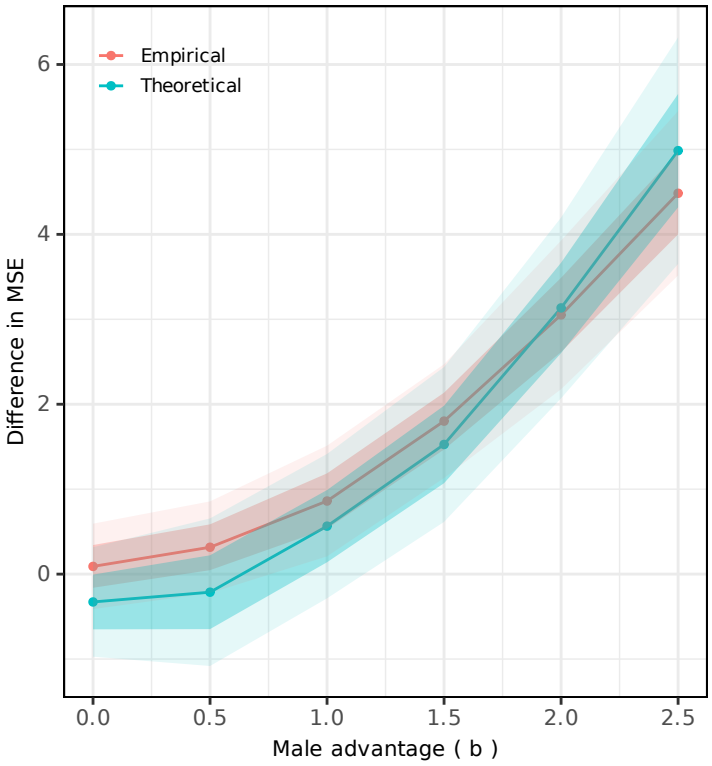

Figure A3: *Theoretical predicted and empirical values of the gap in MSE between a kitchen sink and rubric embeddings model.*

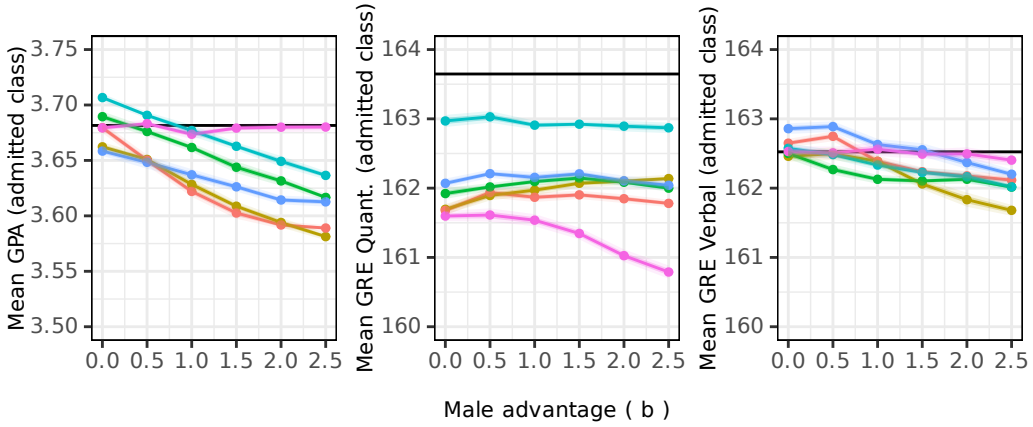

Figure A4: *The impact on various measures of the strength of the admitted class using models of the proxy label with the corresponding level b. Colors indicate the bias mitigation technique. Dark and light shading show ± 1 and ± 2 standard errors, respectively. Solid lines show corresponding values 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.*

D Robustness checks

To confirm theory and empirical results hold outside of the gender bias case we study, here we conduct several robustness checks in which we vary several dimensions of our analysis. Overall, we find qualitatively similar results across each dimension of variation. Below, we provide further details regarding each of these robustness checks. Unless noted otherwise, each analysis follows the procedure of the main text: we inject advantage $b \in [0, 2.5]$ into the labels, fit models using 10×10-fold nested cross-validation, and admit the top 20% of applicants by predicted score.

D.1 Modeling choices

To assess whether our results are sensitive to different modeling choices, including statistical model family, choice of embedding model, choice of outcome or subset of rubric features, we reproduce our main text results with a variety of alternative specifications, detailed below in order.

Model families To assess whether our results generalize beyond ridge regression, we additionally fit LASSO models and random forests. The top two rows of Fig. A5 demonstrate our results are not specific to ridge regression. Both LASSO and random forest models follow similar patterns to our main ridge results: rubric-embedding models do not track the injected audit bias, while models including general purpose embeddings do. Correspondingly, the quality of the admitted class remains much stronger under decisions made by the rubric embedding models. Notably the empirical results of our random forest models do not track the theoretical lines the same way ridge or LASSO models do, though this behavior is expected as random forests are not perfectly linear in their labels.

Fig. A6 reproduces Fig. 3 for LASSO (top row) and random forest models (bottom row), respectively. At low levels of injected bias, rubric-based models remain competitive in predicting the proxy label. As bias increases, black-box and kitchen-sink models fit the proxy label more closely, while rubric-based models maintain lower RMSE against the original admissions scores.

LASSO models were fit with the same computational setup as the ridge models, though we allocated 16 cores instead of 4, decreasing compute time to 3 hours.

Random forests were fit on a CPU cluster using R 4.4 and the `ranger` R package [54]. Per tuning task, we allocated 4 CPU cores, 32 GB memory, and up to 12 hours of wall time on a remote compute cluster. However, the trees fit without exceeding memory usage of 2 GB, and all jobs completed within less than 4 hours. We fit 500-tree forests with 10 outer folds and 10-fold inner calibration. The main grid comprised 720 tasks (3 feature sets x 2 document variants x 6 bias values x 20 random iterations). We selected hyperparameters separately for each feature-set/document-variant/bias specification by five-fold cross-validation over 18 candidate configurations (`mtry` of 10%, 33%, or 67% of features; `min.node.size` of 1, 5, or 10; and `sample.fraction` of 0.632 or 0.8.) Hyperparameters were held fixed across the 20 simulation iterations. The AUC analysis comprised 120 analogous tasks. Based on the observed task durations, we estimate that the complete RF suite required on the order of 2,000–3,000 CPU-hours, however this is an imputed estimate rather than a scheduler-accounting total.

Other embedding models Throughout our analysis, the general-purpose embeddings we used were from OpenAI's `text-embedding-3-large` model [32]. To confirm our results do not depend on this choice, we replicate our main text results using the open source `mxbai-embed-large-v1` embedding model [55], generated locally using the `transformers` package [56]. Similar to `text-embedding-3-large`, we find models fit on `mxbai-embed-large-v1` achieve a high AUC in predicting gender ($98.2\% \pm 0.7\%$), and that $\mathbb{E}[\pi(T(X, m)) - \pi(T(X, f))]$ (the extent to which our representations encode gender), is $0.67 \pm .02$. This corresponds to the smaller slope of the dashed lines we see in Fig. A5, bottom row, as our theoretical results predict. Despite the embeddings' slightly weaker encoding of gender, our audit bias and class composition results persist.

Alternative outcome Our primary analysis treats observed admissions scores as a fixed reference outcome, although these scores may themselves contain bias. We therefore repeat the analysis using first-year core quantitative GPA among students who matriculated in 2024, an outcome much less likely to be subject to bias. We construct representations from their application materials and inject bias by inflating male and deflating female GPAs, varying b from 0 to 1 (and noise standard deviation

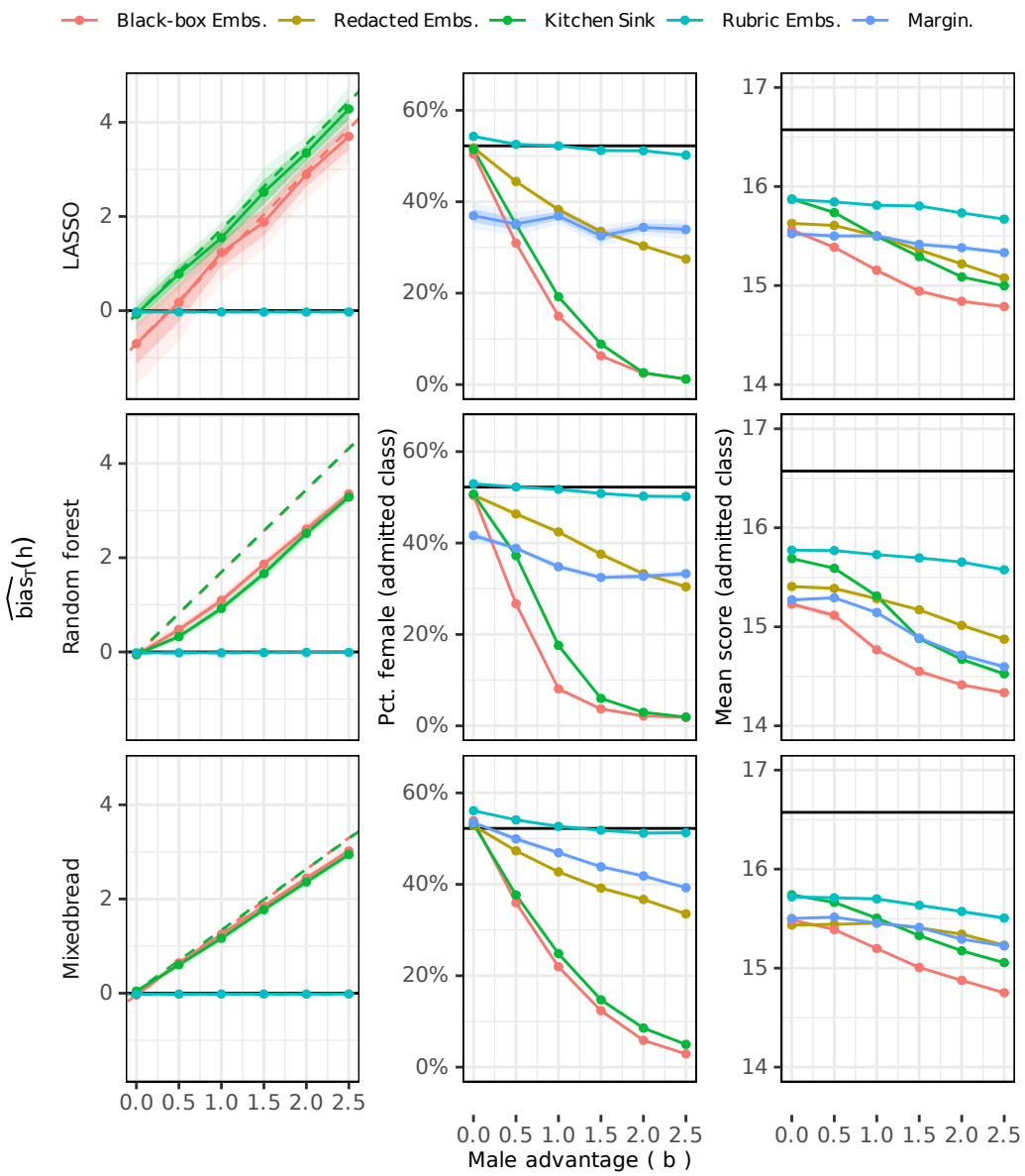

Figure A5: *Replication of Fig. 2 for LASSO models (top row), random forests (middle row), and ridge models fit on embeddings from* `mxbai-embed-large-v1` *[55] (bottom row).* Left: *Estimated bias* $\widehat{\text{bias}}_T(h)$ *of fitted models* $h(x)$*. Dashed line shows bias estimated according to Eq.* (3). Middle: *Percentage of women in the admitted class under a top-20% policy.* Right: *Mean true score* Y *of the admitted class. Shading shows* ± 1 *and* ± 2 *standard errors, respectively.*

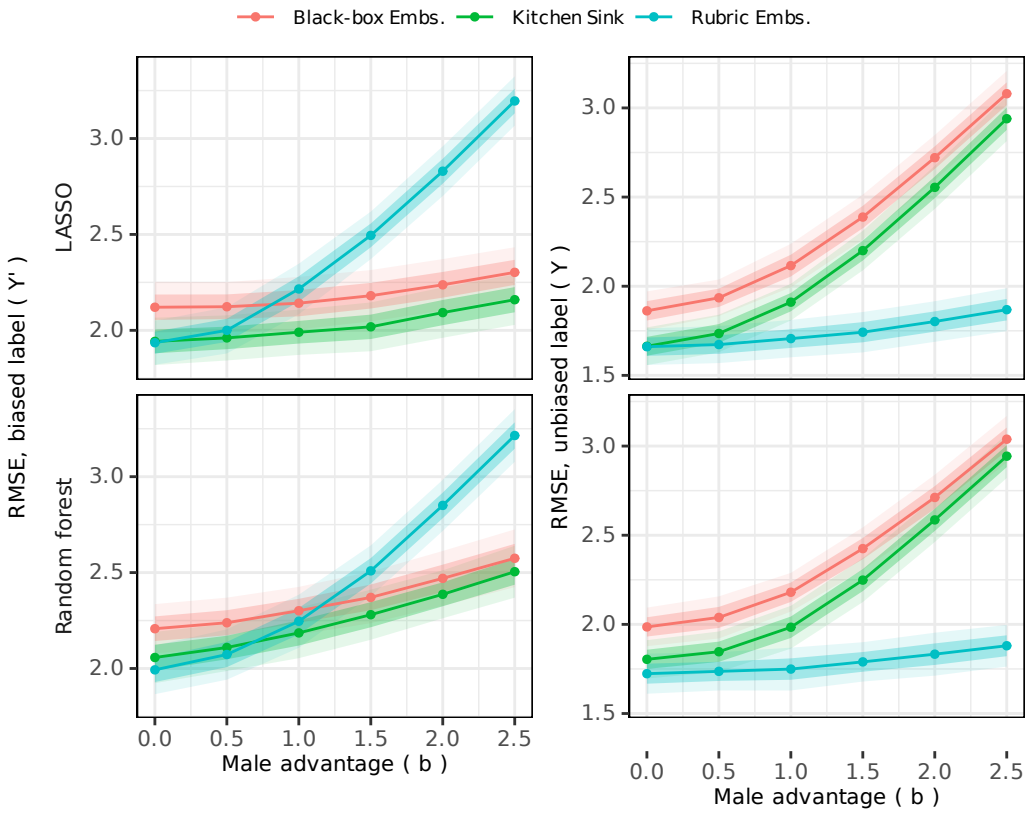

Figure A6: *Replication of Fig. 3 for LASSO models (top row) and random forests (bottom row).* Left: *RMSEs of rubric embedding, black-box embedding, and kitchen sink models, evaluated relative to proxy labels Y'.* Right: *RMSEs of rubric embedding, black-box embedding, and kitchen sink models, evaluated relative to ground truth Y. Shading shows ± 1 and ± 2 standard errors, respectively. At $b = 0$, proxy labels retain zero-mean, unit-variance noise, so RMSEs against Y' and Y need not coincide.*

of 0.5). Within this matriculant sample, we select the top 20% by predicted GPA. Rubric-based models select students with higher observed GPAs than black-box models and show substantially less change in selected-cohort gender composition as the injected advantage increases (Fig. A7, bottom row). We do not estimate audit bias because we did not generate gender-transformed materials for this cohort.

Fewer rubric features To assess if the success of rubric embeddings is dependent on certain features we fit models on random subsets of k of the 811 total features included in our rubric embedding models (396 rubric features, structured covariates, and missingness indicators), from $k = 81$ (10%) to the full set. Bias remains near zero for every k, but the quality of the admitted class increases with k (Fig. A8, top row). We find a similar pattern for GPA and GRE scores (bottom row). Bias stays low even with few rubric items, though models with the most features admit the highest scoring class.

D.2 Forms of bias

To explore whether our results are dependent on our choice of simulated bias, we reproduce our results two ways: we add an explicit bias on race instead of gender (an alternative form of disparate treatment) and a bias towards features that are correlated with being male (to test disparate impact).

Race To assess if our results replicate under a bias for race, as opposed to gender, we create biased labels Y_i' in an identical process to that presented in (1), however instead where $G_i \in \{w, nw\}$

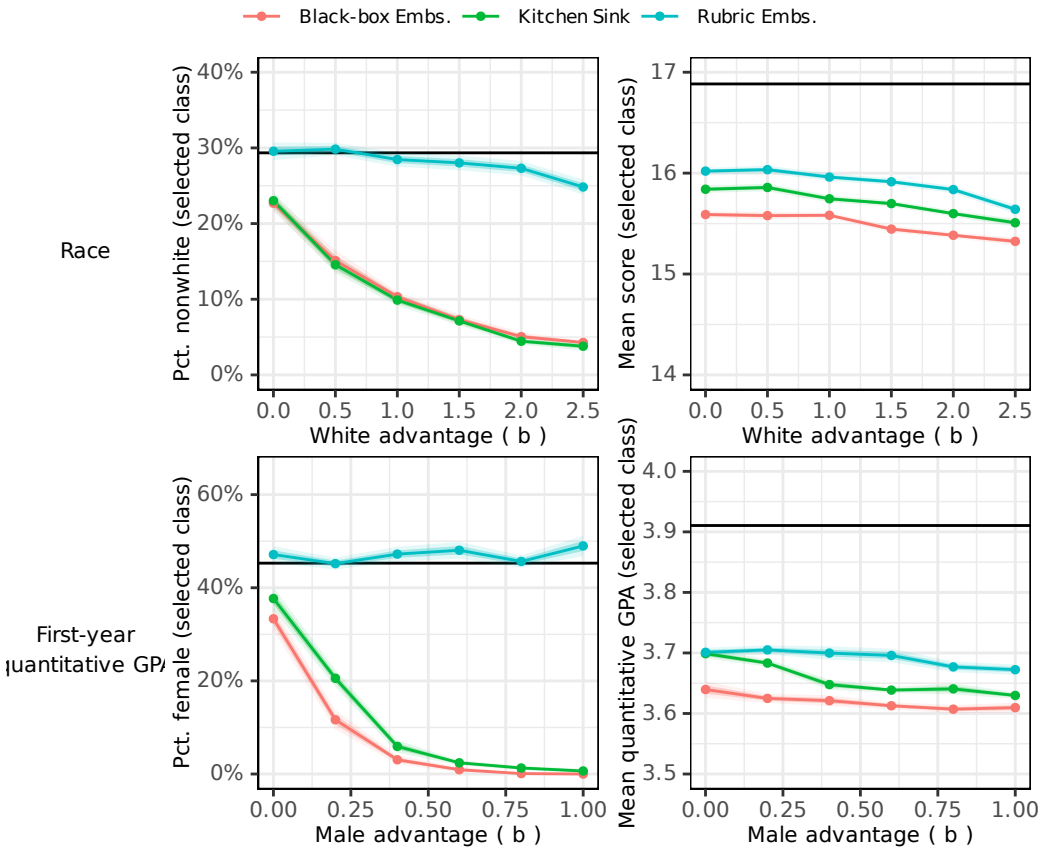

Figure A7: *Replication of Fig. 2, omitting the audit bias panel. The bottom row shows our results when advantage is applied to observed outcomes once admitted (GPA). The top row displays our results when advantage is given to white applicants (and non-whites are disadvantaged).* Left: *Percentage of non-whites (top) or women (bottom) in the admitted class under a top-20% policy.* Right: *Mean true score Y of the admitted class (top), mean quantitative GPA of the admitted class (bottom).*

denotes whether the applicant is white or non-white and b gives the size of the advantage for white applicants. We did not generate the counterfactual materials $T(x, w)$ and $T(x, nw)$ representing the "white-presenting" and "non-white-presenting" versions of an applicant's materials. Without race-based audit study materials, it is not possible to carry out our empirical audit study analysis. Additionally, this analysis is restricted to domestic applicants, as an overwhelming majority of international applicants are missing race information. Regardless, as in the case of gender-biased labels, we see that predictive models trained on "race-biased" labels using black-box embeddings replicate the label bias in their predictions, whereas models trained using rubric embeddings fail to replicate the bias. Further, the class-composition results presented in the top row of Fig. A7 suggest the rubric-embedding based models select cohorts with higher original admissions scores and substantially less change in racial composition than black-box or kitchen-sink models. Consistent with this pattern, we find race is also encoded in general purpose embeddings: models trained to predict race using general purpose embeddings achieve an AUC of $91.9\% \pm 3.0\%$.

Variables correlated with gender We also inject an advantage for business, economics, and engineering majors, who disproportionately identify as male in our sample, and penalize other majors. Fig. A9 reproduces our main results under these biased labels. To avoid shrinking the coefficients on the rewarded major indicators, we exempt these three features from regularization in models that include them. Rubric-based models retain near-zero estimated audit bias and select a greater proportion of women than black-box models. However, their selected cohorts' original admissions

scores decline as the advantage increases, eventually falling below those selected by black-box models. Because the rewarded major indicators are directly represented in the rubric, restricting the representation does not prevent the model from learning distorted weights on these retained features. This experiment illustrates a limitation of rubric embeddings: low audit bias does not ensure protection against all forms of label bias.

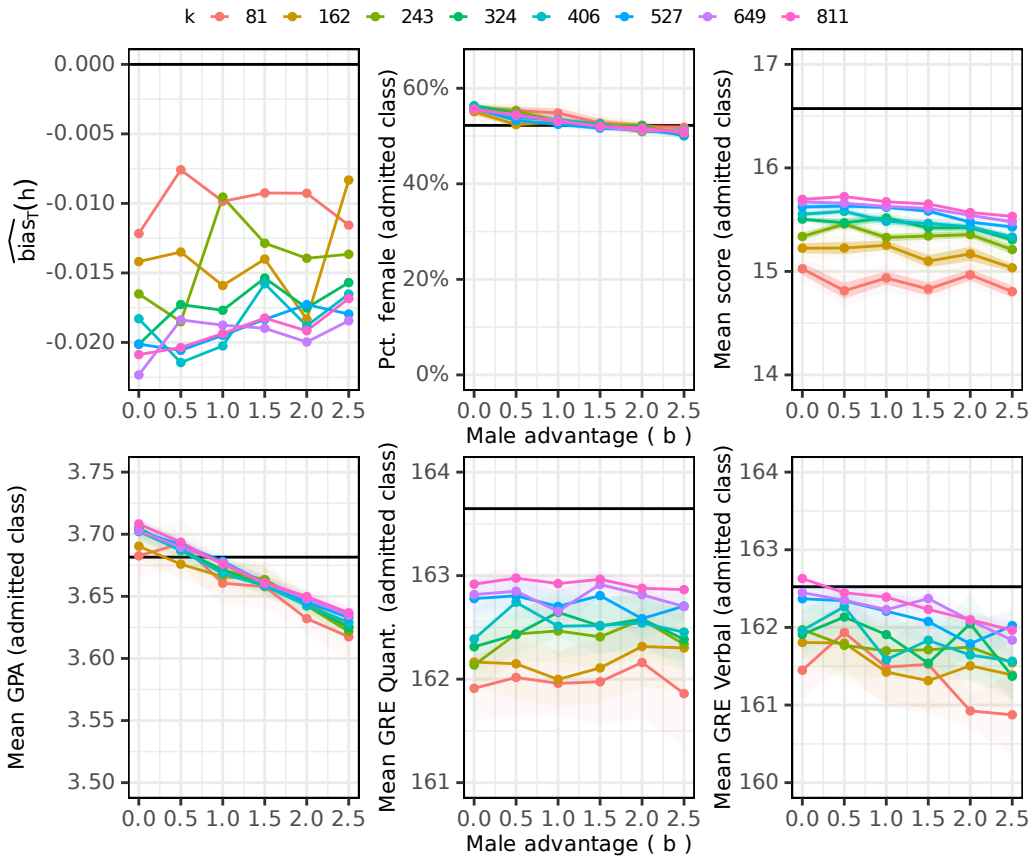

Figure A8: *Rubric embedding models fit on random subsets of k features. Panels as in Fig. 2 and Fig. A4. Shading shows ± 1 and ± 2 standard errors, respectively. Error bars are omitted from the top left panel for visual clarity, all overlap substantially.*

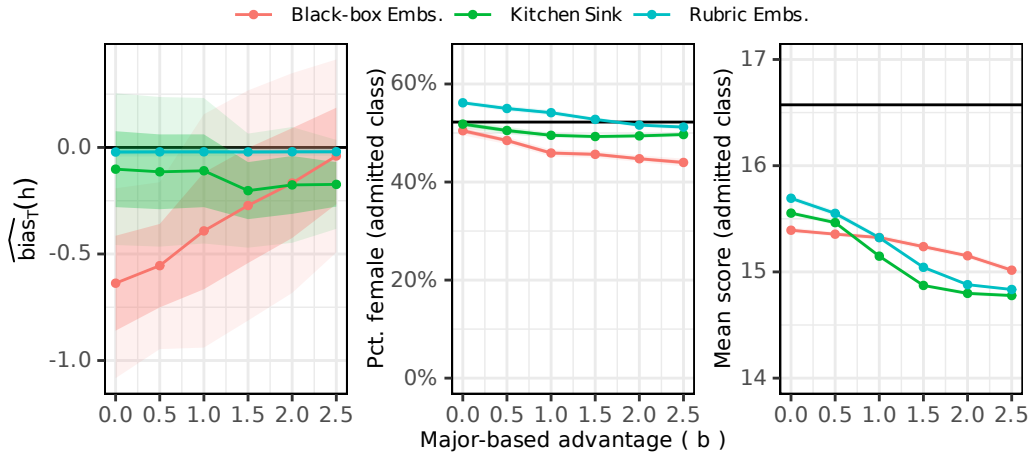

Figure A9: *Replication of Fig. 2 with an advantage for business, economics, and engineering majors, rather than for men.*

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.*

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In Section 3.2, we provide empirical evidence supporting our claim that rubric embeddings mitigate label bias, and demonstrate that our empirical findings align with the theoretical results presented in Section 2.1. Figure 2 demonstrates that our approach indeed reduces group disparities while improving measures of cohort quality. Section 2 and Figure 1 support our claim that training models on proxy labels—particularly models trained on black-box text embeddings—can reproduce biases present in historical evaluations. Our theoretical results (Proposition 2) and the subsequently reported AUCs provide support of our claim that models trained on black-box embeddings are guaranteed to replicate biases in the training data. Figures 2 and A4 support our claims of the strength and benefits of rubric embeddings.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 4 discusses rubric construction, limitations in addressing different forms of bias, restrictions on using group-specific information, and generalizability.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: For proof of all theoretical results presented in the main text, see Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We release our analysis and plotting scripts along with the model outputs necessary to reproduce the figures and inline statistics reported in the paper. Raw application data, extracted feature sets, and explicit rubrics used in feature extraction cannot be released due to FERPA and institutional data agreements. We provide information about our feature extraction process in the Appendix (see Section B). Further, we provide sufficient detail about our process for generating audit study materials in Section C for replication of our approach on new data. The replication package includes the full implementation of all statistical methods in `R/utils.R` and `models.R`.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Raw application data, extracted features, and rubrics used in the extraction process cannot be released due to FERPA and institutional data agreements with the program we study. However, we do release the analysis and plotting scripts (`models.R` and `clean_analysis.R`), all model outputs needed to reproduce each figure in the paper, the full implementation of statistical methods (`R/utils.R`, `R/*_forest_*.R`, and `models.R`, with data-relevant functions stubbed-out and documented) and the audit study prompts used in the correspondence experiment, as well as a detailed description of the pipeline used (Section C). The replication package includes a short `README` describing the contents and provides instructions for reproducing the results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Full experimental details are provided in Section 2 and implemented in `models.R` and `R/utils.R`. Hyper parameter selection for LASSO and ridge models is implemented in `R/utils.R`. We detail our hyperparameter selection for random forests in Appendix D.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All figures report pointwise approximate 68% and 95% confidence intervals as described in figure captions. All inline statistics similarly report approximate 95% confidence intervals. Standard error calculations are implemented in `clean_analysis.R`.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Compute resources for model fitting are described in Section 2 and D. All experiments are CPU-only. OpenAI embeddings and structured feature extraction are conducted through the OpenAI API. Embeddings from `mxbai-embed-large-v1` [55] were obtained locally using the `transformers` Python package [56].

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our data uses sensitive admissions data under an institution data agreement that governs its use and storage. The analysis we conduct and results we share both fall within the scope of that data use agreement. Our work directly addresses bias and fairness concerns in decision making, consistent with the guidelines specified in the code of conduct.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Section 4.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We release analysis code, audit prompts, and aggregate experimental outputs, but do not release applicant materials, applicant-level features, or trained admissions models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, all existing software packages used are properly credited. See `glmnet` [35], OpenAI's `text-embedding-3-large` [32], and `gpt-5-mini` [48], Microsoft Document Intelligence [31], `ranger` [54], `mixedbread-ai/mxbai-embed-large-v1` [55], and the `transformers` package [56]. The commercial APIs used in the paper are used in accordance with their terms of service.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The replication package documents the analysis code and model outputs in a README, including reproduction instructions and restrictions on access to the underlying application data and features. Prompts are available in Prompt A1–Prompt A3.

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: This paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs are a core component of our methodology, used for rubric feature extraction, audit study template generation, validation and repair. We document how we use LLMs extensively; see Sections 3.2 and C.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.